

Information Visualization for Decision Making : Identifying Biases and Moving Beyond the Visual Analysis Paradigm

Evanthia Dimara

▶ **To cite this version:**

Evanthia Dimara. Information Visualization for Decision Making : Identifying Biases and Moving Beyond the Visual Analysis Paradigm. Human-Computer Interaction [cs.HC]. Université Paris Saclay (COmUE), 2017. English. ⟨NNT : 2017SACLS367⟩. ⟨tel-01968033v2⟩

HAL Id: tel-01968033

https://theses.hal.science/tel-01968033v2

Submitted on 7 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

université PARIS-SACLAY

NNT: 2017SACLE367

Thèse de doctorat

Information Visualization for Decision Making

Identifying Biases & Moving Beyond the Visual Analysis Paradigm

Thèse de doctorat de l'Université Paris-Saclay
préparée à Université Paris Sud

École doctorale n°580
Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Gif-Sur-Yvette, le 30 Novembre 2017

Evanthia Dimara

Composition du jury:

Pierre Dragicevic,	Chargé de recherche, Inria Saclay	Directeur de thèse
Anastasia Bezerianos,	Maître de conférences, Université Paris Sud	Co-directeur de thèse
Jo Wood,	Professeur, City, Université de Londres	Rapporteur
Steven Franconeri,	Professeur, Université Northwestern	Rapporteur
Margit Pohl,	Professeur, Université Technique de Vienne	Examinatrice
Christian Jacquemin,	Professeur, Université Paris Sud	Président

Abstract

There are problems neither humans nor computers can solve alone. Computer-supported visualizations are a well-known solution when humans need to reason based on a large amount of data. The more effective a visualization, the more complex the problems that can be solved. In information visualization research, to be considered effective, a visualization typically needs to support data comprehension. Evaluation methods focus on whether users indeed understand the displayed data, can gain insights and are able to perform a set of analytic tasks, e.g., to identify if two variables are correlated.

This dissertation suggests moving beyond this "visual analysis paradigm" by extending research focus to another type of task: decision making. Decision tasks are essential to everybody, from the manager of a company who needs to routinely make risky decisions to an ordinary person who wants to choose a career life path or simply find a camera to buy. Yet decisions do not merely involve information understanding and are difficult to study. Decision tasks can involve subjective preferences, do not always have a clear ground truth, and they often depend on external knowledge which may not be part of the displayed dataset. Nevertheless, decision tasks are neither part of visualization task taxonomies nor formally defined. Moreover, visualization research lacks metrics, methodologies and empirical works that validate the effectiveness of visualizations in supporting a decision.

This dissertation provides an operational definition for a particular class of decision tasks and reports a systematic analysis to investigate the extent to which existing multidimensional visualizations are compatible with such tasks. It further reports on the first empirical comparison of multidimensional visualizations for their ability to support decisions and outlines a methodology and metrics to assess decision accuracy. It further explores the role of instructions in both decision tasks and equivalent analytic tasks, and identifies differences in accuracy between those tasks.

Similarly to vision science that informs visualization researchers and practitioners on the limitations of human vision, moving beyond the visual analysis paradigm would mean acknowledging the limitations of human reasoning. This dissertation reviews decision theory to understand how humans should, could and do make decisions and formulates a new taxonomy of cognitive biases based on the user task where such biases occur. It further empirically shows that cognitive biases can be present even when information is well-visualized, and that a decision can be “correct” yet irrational, in the sense that people’s decisions are influenced by irrelevant information.

This dissertation finally examines how biases can be alleviated. Current methods for improving human reasoning often involve extensive training on abstract principles and procedures that often

appear ineffective. Yet visualizations have an ace up their sleeve: visualization designers can re-design the environment to alter the way people process the data. This dissertation revisits decision theory to identify possible design solutions. It further empirically demonstrates that enriching a visualization with interactions that facilitate alternative decision strategies can yield more rational decisions.

Through empirical studies, this dissertation suggests that the visual analysis paradigm cannot fully address the challenges of visualization-supported decision making, but that moving beyond can contribute to making visualization a powerful decision support tool.

Keywords: information visualization, decision making, cognitive biases, attraction effect, human-computer interaction, behavioral economics

Abstract in French

Certains problèmes ne peuvent être résolus ni par les ordinateurs seuls ni par les humains seuls. La visualisation d'information est une solution commune quand il est nécessaire de raisonner sur de grandes quantités de données. Plus une visualisation est efficace, plus il est possible de résoudre des problèmes complexes. Dans la recherche en visualisation d'information, une visualisation est généralement considérée comme efficace quand elle permet de comprendre les données. Les méthodes d'évaluation cherchent à déterminer si les utilisateurs comprennent les données affichées et sont capables d'effectuer des tâches analytiques comme, par exemple, identifier si deux variables sont corrélées.

Cette thèse suggère d'aller au-delà de ce "paradigme de l'analyse visuelle" et élargir le champ de recherche à un autre type de tâche: la prise de décision. Les tâches de décision sont essentielles à tous, du directeur d'entreprise qui doit prendre des décisions importantes à l'individu ordinaire qui choisit un plan de carrière ou désire simplement acheter un appareil photo. Néanmoins, les décisions ne se résument pas à la simple compréhension de l'information et sont difficiles à étudier. Elles peuvent impliquer des préférences subjectives, n'ont pas toujours de vérité de terrain, et dépendent souvent de connaissances externes aux données visualisées. Pourtant, les tâches de décision ne font pas partie des taxonomies de tâches en visualisation et n'ont pas été bien définies. De plus, la recherche manque de métriques, de méthodes et de travaux empiriques pour valider l'efficacité des visualisations pour la prise de décision.

Cette thèse offre une définition opérationnelle pour une classe particulière de tâches de décision, et présente une analyse systématique qui identifie les visualisations multidimensionnelles compatibles avec ces tâches. Elle présente en outre la première comparaison empirique de techniques de visualisation multidimensionnelle basée sur leur capacité à aider la décision, et esquisse une méthodologie et des métriques pour évaluer la qualité des décisions. Elle explore ensuite le rôle des instructions dans les tâches de décision et des tâches analytiques équivalentes, et identifie des différences de performance entre les deux tâches.

De même que les sciences de la vision informent la visualisation d'information sur les limites de la vision humaine, aller au-delà du paradigme de l'analyse visuelle implique de prendre en compte les limites du raisonnement humain. Cette thèse passe en revue la théorie de la décision afin de mieux comprendre comment les humains prennent des décisions, et formule une nouvelle taxonomie de biais cognitifs basée sur la tâche utilisateur. En outre, elle démontre empiriquement que des biais peuvent être présents même quand l'information est bien visualisée, et qu'une décision peut être "correcte"

mais néanmoins irrationnelle, dans le sens où elle est influencée par des informations non pertinentes.

Cette thèse examine finalement comment mitiger les biais. Les méthodes pour améliorer le raisonnement humain reposent souvent sur un entraînement intensif à des principes et à des procédures abstraits, qui se révèlent souvent peu efficaces. Les visualisations offrent une opportunité dans la mesure où ses concepteurs peuvent remodeler l'environnement pour changer la façon dont les utilisateurs assimilent les données. Cette thèse passe en revue la théorie de la décision pour identifier de possibles solutions de conception. De plus, elle démontre empiriquement que supplémenter une visualisation par des interactions qui facilitent des stratégies de décision alternatives peut mener à des décisions plus rationnelles.

Via des études empiriques, cette thèse suggère que le paradigme de l'analyse visuelle n'est pas en mesure de relever tous les défis de la prise de décision aidée de la visualisation, mais qu'aller au-delà peut contribuer à faire de la visualisation un puissant outil de prise de décision.

Mots clés en français: visualisation d'information, prise de décision, biais cognitifs, effet d'attraction, interaction homme-machine, économie comportementale

Dedication and acknowledgements

This section is the oscar speech of an otherwise strictly scientific manuscript.

I sincerely wish to all Ph.D. students mentors like Pierre Dragicevic and Anastasia Bezerianos. Pierre, thank you for making me a researcher, first by teaching me how to develop and defend my ideas. I truly believe that you are of the most intelligent and kind debaters I have ever met. Thank you for showing me how to be honest and transparent with my findings and how to be pedagogical when communicating them. I am particularly grateful for these two: that you were bold enough when I insisted to express provocative ideas, but at the same time, you were spending a fair amount of time helping me to polish and support them. Anastasia, thank you for teaching me how to design rigorous user studies, how to write a scientific article, how to focus on the important, how to cooperate with people - and most importantly, for being so open-minded giving me the space to develop my own ideas. I will never forget from both of you your sincere emotional support, your caring and your thoughtful advices on my future career choices.

I very thank the members of my jury: Steven Franconeri, Jo Wood, Margit Pohl, Christian Jacquemin for their insightful and so generous reports, and their inspiring questions and feedback the day of the defense. Steven, I feel I own you special thanks for your report. Your incredibly generous comments will encourage me in every future career step.

I would like to also express my gratitude to Digiteo for accepting to fund my own Ph.D. proposal allowing me to remain constantly excited about my research topic.

I would also like to thank all members of the Aviz and ILDA teams (alumni and newcomers) for offering me their help in every practical issue I had, their precious feedback on my research, for attending my unpolished talks and for making the conferences so enjoyable.

I want to thank my parents among many others, for these three: their unconditional love, their (non-objective) belief on my abilities and their generosity to make me an independent person. Mom and dad thank you for being always my hidden power source.

Although an only-child, I am lucky enough to have a sister and a brother of whom their love and support helped me cope with these three years of research. Eirini Glynou, thank you for feeding the revolution monkey and helping me to place an otherwise geeky computer science thesis to things that matter. Emmanouil Giannisakis, thank you for being the kind of friend that when you wake him up three o'clock in the morning, instead of throwing you a pillow, he makes you a tea with sympathy and patiently elaborates on your research failures.

Having said that, there many ways of finishing a Ph.D. My way happened to be with a devil giggle, yet with confidence and a feeling of fulfillment. The reason is Achille Tuglo. Achille, thank you for being always there, protecting me with your love from the hungry beast of academia.

Table of Contents

Page

List of Tables **xv**

List of Figures **xvii**

1 Introduction **1**

1.1 Thesis Statement . . . 8

1.2 Thesis Overview . . . 10

2 Background **13**

2.1 Visualization Tasks . . . 14

2.1.1 Analytic tasks . . . 15

2.1.2 Decision task . . . 17

2.1.3 Summary . . . 20

2.2 Decision Theory . . . 20

2.2.1 Normative decision models . . . 20

2.2.2 Descriptive decision models . . . 21

2.2.3 Summary . . . 24

2.3 Cognitive Biases . . . 24

2.3.1 What is a cognitive bias . . . 24

2.3.2 Criticism of cognitive biases . . . 26

2.3.3 Taxonomies of cognitive biases . . . 26

2.3.4 FAULTY: A new task-based taxonomy of cognitive biases . . . 28

2.3.5 Biases of a faulty choice task . . . 31

2.3.6 Biases of a faulty estimation task . . . 33

2.3.7 Biases of a faulty recall task . . . 34

2.3.8 Biases of a faulty hypothesis assessment task . . . 34

2.3.9 Biases of a faulty attribution task . . . 35

2.3.10 Other categories . . . 36

2.3.11 Conclusion . . . 37

2.4 Decision Support Visualization Systems . . . 38

2.4.1 Overview . . . 38
2.4.2 Evaluation of visualizations for decision support . . . 42
2.5 Conclusion . . . 44

3 Multidimensional Visualizations for Decision Making 47
3.1 Multidimensional Visualization Approaches . . . 48
3.1.1 Dimensionality reduction . . . 48
3.1.2 Non-Geometric visualization techniques . . . 49
3.1.3 Lossless geometric projection . . . 50
3.1.4 Evaluations . . . 52
3.2 Technique Design . . . 52
3.2.1 Visual encodings . . . 53
3.2.2 Interaction techniques . . . 53
3.3 Experiment . . . 54
3.3.1 Research questions . . . 56
3.3.2 Tasks . . . 56
3.3.3 Datasets and task generation . . . 57
3.3.4 Apparatus . . . 57
3.3.5 Techniques . . . 58
3.3.6 Participants . . . 58
3.3.7 Experiment design . . . 58
3.3.8 Procedure . . . 58
3.3.9 Objective performance metrics . . . 59
3.3.10 Subjective metrics . . . 61
3.4 Results . . . 62
3.4.1 Planned analyses . . . 62
3.4.2 Additional analyses . . . 66
3.5 Discussion . . . 69
3.6 Conclusion . . . 70

4 The Role of Instructions in Decision Making Tasks 73
4.1 Narratives in Information Visualization . . . 74
4.1.1 Question wording . . . 75
4.1.2 Visualization narratives . . . 75
4.1.3 Illustrative use cases . . . 76
4.1.4 Narratives in visualization evaluation . . . 76
4.2 Narratives in crowdsourcing . . . 77
4.3 Experiment . . . 78
4.3.1 Dataset and visualization . . . 78

4.3.2 Tasks . . . 79
4.3.3 Context conditions . . . 80
4.3.4 Objective performance metrics . . . 81
4.3.5 Subjective metrics . . . 82
4.3.6 Experiment design . . . 82
4.3.7 Research questions . . . 83
4.3.8 Overview of results . . . 84
4.3.9 Planned analyses . . . 89
4.4 Discussion . . . 91
4.4.1 On the use of narratives in experiment instructions . . . 92
4.4.2 On decision making narratives . . . 93
4.5 Conclusion . . . 93

5 Detecting Cognitive Biases in Visualization Systems 95
5.1 The Attraction Effect . . . 97
5.1.1 Why does the attraction effect occur? . . . 98
5.1.2 Can the attraction effect occur with visualizations? . . . 98
5.2 Gym Experiment: Table/Scatterplot, 3 Choices . . . 100
5.2.1 Design rationale . . . 101
5.2.2 Experiment design . . . 105
5.2.3 Results . . . 107
5.3 Extending the Attraction Effect . . . 111
5.3.1 Can the attraction effect generalize to many alternatives? . . . 111
5.3.2 Ways of adding more alternatives . . . 112
5.3.3 Extended procedure . . . 112
5.4 Real Experiment: Scatterplot, Many Choices, Realistic dataset . . . 113
5.4.1 Design rationale . . . 113
5.4.2 Terminology . . . 114
5.4.3 Stimuli and task . . . 114
5.4.4 Scenario and attribute values . . . 115
5.4.5 Measures . . . 118
5.4.6 Crowdsourcing quality control . . . 118
5.4.7 Experiment design . . . 119
5.4.8 Participants . . . 120
5.4.9 Procedure . . . 120
5.4.10 Hypothesis . . . 120
5.4.11 Results . . . 120
5.4.12 Planned analyses . . . 120
5.4.13 Additional analyses . . . 121

5.4.14 Discussion . . . 121
5.5 Bet Experiment: Scatterplot, Many Choices . . . 122
5.5.1 Design rationale . . . 122
5.5.2 Experiment design . . . 126
5.5.3 Results . . . 128
5.6 Discussion . . . 130
5.7 Conclusion . . . 131

6 Towards Improving Decision Support Visualization Systems 133
6.1 DcPairs: Multi-attribute Decision Support with Annotations . . . 133
6.1.1 The use of color in decision support visualizations . . . 133
6.1.2 Attributes encoding in decision support . . . 136
6.1.3 The use of scatterplot matrices in decision support . . . 136
6.1.4 The use of personal annotations in visualizations . . . 137
6.1.5 Communication of uncertainty . . . 139
6.1.6 Attribute scaling . . . 141
6.1.7 DcPairs use case . . . 142
6.1.8 Summary . . . 144
6.2 Debiasing Techniques . . . 145
6.2.1 Educational debiasing techniques . . . 145
6.2.2 Motivational debiasing techniques . . . 146
6.2.3 Computation-aided debiasing techniques . . . 147
6.2.4 Group-based debiasing techniques . . . 147
6.2.5 Design-based debiasing techniques . . . 148
6.2.6 Summary . . . 148
6.3 Deletion Experiment: Alleviation of the Attraction Effect . . . 149
6.3.1 Design rationale . . . 149
6.3.2 The deletion task . . . 151
6.3.3 Experiment task . . . 152
6.3.4 Interactions . . . 153
6.3.5 Experiment design . . . 153
6.3.6 Participants . . . 153
6.3.7 Procedure . . . 155
6.3.8 Hypotheses . . . 155
6.3.9 Planned analysis results . . . 155
6.3.10 Additional results . . . 156
6.3.11 Discussion . . . 156
6.4 Conclusion . . . 158

7 General Conclusion & Perspectives **159**
7.1 Summary of Lessons Learned . . . 159
7.2 Summary of Limitations . . . 164
7.3 Summary of contributions . . . 168
7.4 Thesis Conclusion . . . 170
7.5 Future Perspectives . . . 171
7.5.1 FAULTY: Cognitive bias exploration . . . 171
7.5.2 Group decision making . . . 175
7.5.3 Computation-aided decision making . . . 176
7.5.4 Final words . . . 177
ITHACA . . . 178

A Appendix A **179**
A.1 FAULTY taxonomy of cognitive biases . . . 179

B Appendix B **185**
B.1 Decision-Support Visualization Systems . . . 185

Bibliography **189**

233

List of Tables

Table		Page
4.1	The instructions text in each experiment condition. The condition O-NAR has identical page 1 with AN-NAR and identical page 2 with SEM condition.	85
A.1	Color legend of the task-based categories of cognitive biases	179
A.2	Color legend of the relation of each cognitive bias to visualization research	179
A.3	Alternative names for cognitive biases	183
B.1	Color Legend for Decision Support Visualization Systems	185
B.2	Decision Support Visualization Systems Index	186

List of Figures

Figure **Page**

1.1 U.S. Intelligence Community. (sources: left [8] , right [9]) . . . 2
1.2 Doctor with patient (left). Bank advisor with client (right). (sources:[6],[1]) . . . 2
1.3 Print advertisement created by Leo Burnett, Brazil for AE. (source: [12]) . . . 3
1.4 [LEFT] United States presidential election, 1992, [RIGHT] London local elections, 2010 (source:Wikipedia) . . . 4
1.5 The confirmation bias. (source: chainsawsuit.com) . . . 5
1.6 Data visualization. source:dasaa.in . . . 6
1.7 Examples of visualization systems displaying specialized datasets such as medical [337], geographical [77], financial [534], biology [302], and energy consumption datasets [71]. . 7
1.8 Examples of "personal" or "casual" visualization systems displaying personal datasets, e.g., water consumption [177], social networks[222], email [487], daily activities [48], music streams [49]. . . . 7

2.1 Descriptive models: properties of decision strategies (adapted image from Payne et al 1993 [372]) . . . 22
2.2 Optical illusion.[11] . . . 25
2.3 Overview of the FAULTY taxonomy. For the complete taxonomy, see Appendix A. . . . 30
2.4 Zhang et al. [530] . . . 31
2.5 Wood et al. [516] . . . 32
2.6 Micallef et al. [332] . . . 33
2.7 Confirmation Bias. source:chainsawsuit.com . . . 34
2.8 Observer-expectancy effect. source:quibb.com . . . 35
2.9 Examples of dataset-specific visualization systems that target decision-support: **a**: FinVis [405], **b**: Stratos [35], **c**: Financial Portfolio [415], **d**: Shen et al. [427], **e**: VisIDM [127], **f**: LiteVis [439], **g**: Ovis [233], and **h**: Afzal et al. [15]. . . . 39
2.10 Examples of visualization systems designed for exploratory data analysis that are thought to aid multi-attribute choice due to their support for interactive querying and filtering **a**: ScatterDice [156], and **b**: HomeFinder [513]. . . . 40

2.11 Examples of generic decision-support visualization systems for product comparison: **a**: ManyLists [307], **b**: ProductExplorer [396], **c**: FOCUS [443] **d**: SmartClient [389], and **e**: EZChooser [514]. 41
2.12 Examples of generic decision-support visualization systems that allow users to combine multiple attributes into a single aggregate score: **a**: Value Charts [51], **b**: LineUp [201], **c**: WeightLifter [364], and **d**: CommonGIS [24]. 41
2.13 Examples of generic decision-support visualization systems with alternative designs: **a**: Dust&Magnet [525], **b**: DataContextMap [93], and **c**: AHPtreemaps [33]. 42

3.1 The visualizations evaluated in chapter 3: Parallel Coordinates (PC), Scatterplot Matrix (SM) and Tabular Visualization (TV). 47
3.2 Examples of non-geometric visualization techniques: **a**: Chernoff Faces [297], **b**: StarGlyphs [297], **c**: Pixel-oriented [272], **d**: Hierarchical [428], and **e**: Graph-based [53]. 49
3.3 Examples of geometric visualizations based on a scatterplot: **a**: 3D scatterplots [286], **b**: Star Coordinates [267], **c**: Generalized SM[157], and **d**: HyperSlice SM [231]. 50
3.4 Examples of geometric visualizations based on parallel coordinates: **a**: PC Circular [231], **b**: PC Bundled[257], **c**: PC curved [329], and **d**: PC 3D [257]. 51
3.5 Examples of geometric visualizations based on a tabular visulization: **a**: TableLens [312] and **b**: Bertifier [374]. 51
3.6 A) Single data case “France” highlighted on hovering. B) Range selection on two dimensions. The three data cases at the intersection of the selected ranges (63-72 for “Employment rate", and 340000-40000 for “Personal earnings") are “France”, “Germany” and “Finland”. 52
3.7 Experiment Stimuli for the decision task (“Which holiday package do you choose?”). Dataset of 100 holiday packages. 55
3.8 *Left:* Mean accuracy scores achieved for the three analytic tasks and the decision making (DM) task, using the parallel coordinates plot (PC), the scatterplot matrix (SM), and the tabular visualization (TV). *Right:* Mean differences in accuracy scores between each pair of techniques — a positive value indicates that the left technique outperforms the technique on the right. All error bars are 95% CIs (n=21). 62
3.9 Average time (in seconds) spent on each analytic task for techniques PC, SM and TV. All error bars are 95% CIs (n=21). 64
3.10 Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21). 64
3.11 *Left:* Mean rating for each technique, for the analytic (AN) and for the decision (DM) tasks. *Right:* Mean differences in ratings between each pair of techniques — a positive value indicates a preference for the technique on the left. Error bars are 95% CIs (n=21). 65
3.12 *Left:* Mean rating for each choice assessment metric and each technique. *Right:* Mean differences between each pair of techniques — a positive value indicates a benefit (e.g., higher choice satisfaction) for the technique on the left. All error bars are 95% CIs (n=21). 66

3.13 Average time (in sec) spent on the decision making task for techniques PC, SM and TV. All error bars are 95% CIs (n=21). 67

3.14 Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21). 67

3.15 *Upper:* Mean difference in preference for techniques PC, SM and TV from analytic to decision tasks — positive values indicate that people increased their preference for the technique when used it for a decision task. *Bottom:* Mean absolute difference in preference for a technique PC, SM and TV from analytic to decision tasks — positive values indicate that people changed their preference for the technique when used it for a decision task. All error bars are 95% CIs (n=21). 68

4.1 Storytelling outdoors, Concord Library, 1970 by Paul Ife Horne, appeared in Concord. . . 75

4.2 Crowdsourcing. source:https://pixabay.com/en/crowd-lego-staff-choice-selector-1699137 77

4.3 Stimuli used in each task (Ext, Cor and Com), and in the in-task attention test. Correct answers are annotated in color. Axes were labeled (X, Y) for ABS, and ($size\ m^2$, $price$ ($)) in all other context conditions. The title was $Diagram\ Z : Datapoints$ in ABS, and was $Diagram\ Z : Houses$ in SEM (all tasks) and DM-NAR (Ext, Cor tasks). In all other conditions the title was $Agency\ Z : Houses$. Z was an integer (1, 2, 3, or 4) identifying the scatterplot. 79

4.4 Self-reported demographics of our participants. 84

4.5 Indicative chart showing the correspondence between degree of CI overlap and p-values for independent samples (after [290]). 86

4.6 Accuracy per task and condition. Error bars are 95% CIs. 87

4.7 In-task and post-task attention. Error bars are 95% CIs. 87

4.8 Reported confidence scores. Error bars are 95% CIs. 88

4.9 Reported easiness scores. Error bars are 95% CIs. 88

4.10 Enjoyability and Usefulness 89

4.11 *In gray:* Mean differences in accuracy between DM-NAR and AN-NAR across tasks and for Ext and Cor; *In color:* Mean differences in accuracy, in-task and post-task attention between DM-NAR and AN-NAR. Positive values indicate a benefit for DM-NAR. Error bars are 95%CIs. 89

4.12 Mean differences in accuracy, in-task attention and post-task attention between all NAR conditions combined and SEM. Positive values indicate a benefit for narratives. Error bars are 95%CIs. 90

4.13 Mean differences in accuracy, in-task attention and post-task attention between SEM and ABS. Positive values indicate a benefit for SEM. Error bars are 95%CIs. 91

4.14 Mean differences in accuracy, in-task attention and post-task attention between O-NAR and AN-NAR. Positive values indicate a benefit for O-NAR. Error bars are 95%CIs. 91

5.1 Example of an attraction effect in elections: Bob has an excellent education plan, while Alice is very strong in crime control. The addition of Eve, a candidate similar but slightly inferior to Alice, raises Alice's attractiveness as a candidate. This irrelevant option is called a *decoy*. (Photos Benjamin Miller, FSP Standard License, icons by Ivan Boyco, CC-BY license) 95

5.2 Stimulus of attraction effect experiment with hotel photos by Frederick et al. [174] 99

5.3 Stimulus of attraction effect experiment with television photos by Frederick et al. [174] . 99

5.4 Stimulus of attraction effect experiment with fruit photos by Frederick et al. [174] 99

5.5 Stimulus of attraction effect experiment with gambling tickets by [174]: the probability of the ticket is shown as a "probability wheel" and the prize as a number 100

5.6 Examples of experimental stimuli for the table (a,b) and the scatterplot (c,d) conditions. The left multi-attribute choice task (a,c) has no decoy, while the right choice task (b,d) has a decoy on B. A and B are the incomparable alternatives, and the decoy C is in this example on the target B. Presentation in the experiment: size 250x260 cm, font Helvetica, 15px 103

5.7 Participant demographics for both experiments. 106

5.8 Proportions of participant choices in all three experiments. 108

5.9 Point estimates and 95% confidence intervals for the attraction effects in Malkoc et al. [316], and in our two experiments. 109

5.10 Self-reported confidence and familiarity in all three experiments ("Gyms", "Real" and "Bets"). 110

5.11 A baseline choice task T_0 and two possible test choice tasks: T_A, where A is the target, and T_B, where B is the target. 113

5.12 Cars dataset, decoys on A 115

5.13 Cars dataset, decoys on B 115

5.14 Houses dataset, decoy on A 115

5.15 Houses dataset, decoy on B 115

5.16 The *cars* and *houses* datasets with their Pareto front, in pink. Crosses are discarded outliers. 116

5.17 The house dataset D and three possible subsets (D_0, D_A and D_B) created by randomly removing alternatives in specific regions. 117

5.18 The scatterplot reading test. The answer was provided through a radio button. 119

5.19 The non-dominated alternatives used in our choice tasks. 123

5.20 Experimental stimuli for the two matched choice tasks AC and CA (black-and-white background images), and explanatory annotations (box overlays). See Section 5.3 for the full details. 124

5.21 Contingency tables showing choice pairs for all matched tasks 129

6.1 Bartram et al. [45] 134

6.2 This Figure illustrates the use of color in current visualization systems that target decision support. Most systems use color to encode the identity of attributes such as WeightLifter (a [364]), LineUp (b a screenshot from the online version of the tool using the cereals dataset [10, 201]), STRATOS (c [35]). And others (e [13, 33]), to encode choice alternatives such as FilmFinder (d [16]) and AHPtreemaps. . . . 135
6.3 A Generalized Pairs Plot [157] . . . 136
6.4 GPLOM [246] . . . 136
6.5 Scatterplot Matrix DcPairs . . . 137
6.6 Figure from Click2Annotate prototype by Chen et al. [91] . . . 138
6.7 Figure of Business Intelligence (BI) dashboard by Elias & Bezerianos [153] . . . 139
6.8 Annotation feature in DcPairs . . . 139
6.9 Experiment in section 3.1: Self-reported preferences for the attributes of holiday packages. The 21 participants are encoded with distinct colors. Each row shows an attribute (i.e. archaeological interest, landscape interest, nightlife interest, kids friendly, hotel quality, sports activity, and security level). Six observations (bars) per attribute for each participant (timeline order). Positive-negative values indicate direction of their preference, i.e., whether they prefer the attribute to be high or low. . . . 140
6.10 Attribute encodings in DcPairs . . . 140
6.11 Threshold slider . . . 141
6.12 Attribute Map overview. See Figure 6.13 for a larger image. . . . 141
6.13 Use Case Scenario of DcPairs using QS university world rankings 2013 dataset . . . 142
6.14 DcPairs overview with countries dataset from chapter 3. [LEFT] Initial system view before the user assigns importance weights. Identical gray boxes indicate attributes of equal importance. [RIGHT] Attributes are shown as shaded boxes; on the diagonal for a detailed view (A) others in an attribute map overview (B). Attribute weights are shown in a continuous gray scale ( ) expressing preferences as relative tendencies rather than precise values. The user annotates the countries (dots) using self-authored color-coded labels (C). Green dots show the favorite choices. . . . 144
6.15 Clegg et al. [103] . . . 146
6.16 Group decision making. source:pixabay.com . . . 147
6.17 Illustration of the attraction effect in scatterplot visualizations. The position of the decoys affects the attractiveness of the **target** (A for the left plot, C for the right plot) – reminder from Bets experiment in section 5.5 . . . 149
6.18 Possible techniques to alleviate the attraction effect in scatterplots: remove the dominated points (A), highlight the Pareto front (B), de-emphasize the dominated points (C). . . . 150
6.19 Deletion experiment: the task in selection condition (baseline) . . . 152
6.20 Deletion experiment: the task in deletion condition . . . 152
6.21 Deletion training (part2) . . . 154

6.22 Deletion training (part3) . . . 154
6.23 Deletion experiment: self-reported participant demographics . . . 154
6.24 Point estimates and 95% confidence intervals for the attraction effects in Selection and Deletion conditions. . . . 155
6.25 Deletion experiment: Self-reported confidence . . . 156

7.1 Parallel Coordinates (PC), Scatterplot Matrix (SM) and Tabular Visualization (TV). . . . 160
7.2 A sample of instructions given to participants in narratives experiment (Chapter 4). . . . 161
7.3 The taxonomy "FAULTY' of 139 cognitive biases and the attraction effect (the black dot). 162
7.4 Reminder of DcPairs system described in Chapter 6. [LEFT] Initial system view before the user assigns importance weights. Identical gray boxes indicate attributes of equal importance. [RIGHT] Attributes are shown as shaded boxes; on the diagonal for a detailed view others in an attribute map overview. Attribute weights are shown in a continuous gray scale ( ) expressing uncertainty in user preferences. The user annotates using self-authored color-coded labels . . . 163
7.5 FAULTY: taxonomy of cognitive biases. . . . 172
7.6 Six bias metrics by Wall et al.[493] . . . 173
7.7 Possible extensions to the attraction effect: *compromise effect* (A), where peoples' choices are affected by the middle alternatives, and *phantom effect* (B), by dominant but unavailable alternatives. . . . 174
7.8 Which plot was randomly generated? Answer: the left, source [2] . . . 174
7.9 Group decision making with visualizations. source: A webinar led by Tara Nutley, MEASURE Evaluation; Stacey Berlow, Project . . . 175
7.10 Automation bias source: lskitka.people.uic.edu/styled-7/styled-14/ . . . 176

Chapter 1

Introduction

"The limits of my language are the limits of my mind. All I know is what I have words for."

Ludwig Wittgenstein

We often delegate our decisions to experts. When we need to build a house, an architect decides which designs meet the code and local ordinances, or which materials can guarantee performance standards and safety. When we feel sick, our personal doctor decides which medication is appropriate based on the symptoms and our medical history. When we want to invest in the stock market, a professional broker may suggest where to allocate our assets based on her market analysis. Most of these decisions require expertise and specialized knowledge that we lack. Other decisions, such as choosing a career life path or a political view to support, also require access to specialized knowledge, but are considered too personal to delegate to a third party. Still, the more overwhelmed we are by information, the more we tend to place our hopes on experts to make decisions on our behalf.

Nevertheless, experts' decisions are not necessarily good decisions. The Intelligence Community (IC) of the United States (Figure 1.1) is a vast federation consisting of roughly 100,000 employees and a budget comparable to a nation's gross domestic product [167]. The intelligence experts of IC routinely make national decisions such as whether to draw a weapon or how to combat terrorism [395]; risks that involve human lives. Reyna et al.[395] compared in a study the IC expert analysts and college students on their ability to make risky decisions and found that the experts were more prone to irrational mistakes than the college students. Larrick et al. [294] also examined the people's decisions when microeconomic principles are conflicted with humanitarian values (e.g., exploiting blood from poor Asian populations) and found that people with more training in economics are more



Figure 1.1: U.S. Intelligence Community. (sources: left [8] , right [9])

likely to ignore humanitarian consequences in their decisions. Frank [173] also reports that formal exposure to economics can lead people to focus on maximizing individual benefit often in a conflicted manner with the general good.

Conflicts of interest can be a serious problem when delegating decisions to experts. For example, doctors would ideally like to do the best for their patients, but this may entail losing money [191]. In most countries it is legal for doctors to receive "bribes" by pharmaceutical companies for every new patient they put on their drugs [191], and, also, patients often receive an unnecessary surgery or expensive imaging [191] that are rather profitable for the doctor. Similarly, bank advisers often persuade their clients into investments that are more profitable for the bank than for the clients [191]. Even when visiting a store to buy a new laptop, we often hesitate to trust the technical assistant that her suggestions are tailored to our needs rather than the need of the store to profit from an unnecessary purchase. Besides profit, sometimes there are also legal issues involved. For example, doctors may suggest unnecessary tests, drugs, or surgery, even at the risk of hurting a patient, out of fear of being sued if a disease is overlooked (known as "defensive decision making") [191].

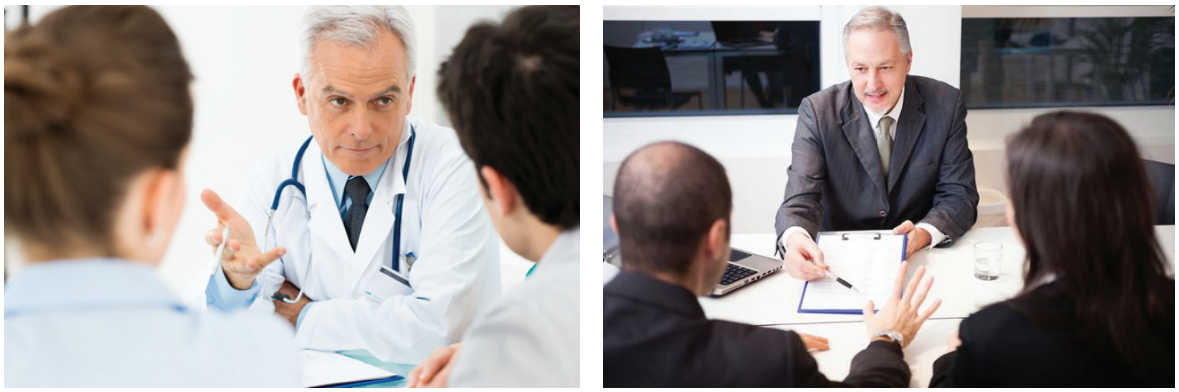


Figure 1.2: Doctor with patient (left). Bank advisor with client (right). (sources:[6],[1])



Figure 1.3: Print advertisement created by Leo Burnett, Brazil for AE. (source: [12])

The question of who should make a decision has always been a subject of controversy in organized societies. Plato argued that philosopher kings should be the rulers, as political rule depends on knowledge, which philosophers possess [537]. Nowadays, decision makers are often selected based on technical expertise to replace elected representatives, known as "technocracy" [409]. For example, during the first years of financial crisis, countries of the European south, particularly Italy and Greece, have seen a change in office with the governments of the technocrats Monti M. and Papademos L. [122, 148, 484]. Applying harsh austerity measures to vulnerable populations [86, 323, 529], the experience of the recent financial crisis has launched a debate on the compatibility of technocracy with democratic values [118, 205, 369].

There are many important decisions that we cannot delegate to experts and that we have to take ourselves. For example, the modern policy in the medical field in the US is that the patients have to decide themselves on the best treatment, even when facing life-threatening diseases. When we listen to the news, we need to form our own views on a topic that can later affect impactful life choices (e.g., our behavior towards the environment). Even in simpler cases, such as the purchase of a new camera, we need to search ourselves a large market to identify the best deal for our needs. Nevertheless, the information involved in most decisions is hard to find and process. As a result, we tend to trust our "gut feelings" or simplify the process by focusing our attention only on small pieces of information. These

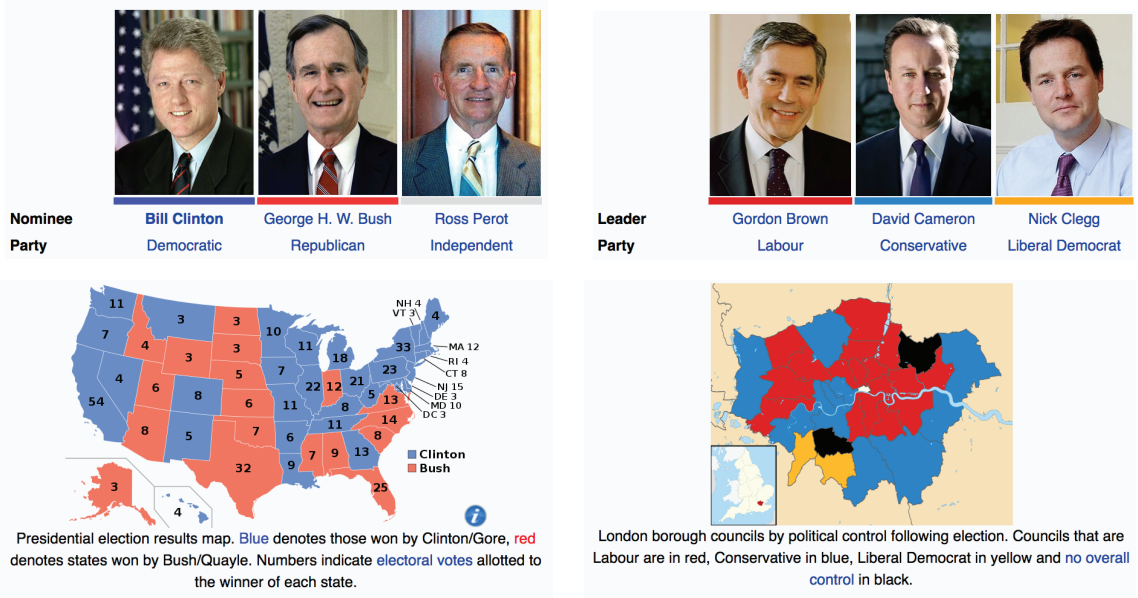


Figure 1.4: [LEFT] United States presidential election, 1992, [RIGHT] London local elections, 2010 (source:Wikipedia)

shortcuts of decision processes can be often wise and save us precious time [190, 191]. In many other cases, though, "gut feelings" can distort our judgments in occasions that we fail to oversee [263, 382].

We want to make decisions that lead to the best outcomes for our lives, but we rarely realize how often the way we evaluate our options is distorted. For example, assume we want to choose between two holiday packages for two appealing destinations, Paris and New York. If a travel agency wants us to choose the New York package, they can apply a simple marketing trick: to add a third holiday package for New York at a higher price. This third package – even though no one will ever choose as it is more expensive– can make the similar, but cheaper New York package look more attractive than the Paris one. This is a well known marketing trick called **attraction effect** (or the decoy effect) and it has been observed in several contexts, even in more critical decisions such as in political elections. In 1992, there was a presidential election race between H. W. Bush and Bill Clinton (Figure 1.4). O'Curry and Pitts showed evidence that the entry of Ross Perot into the race may have led Bill Clinton to be perceived to be more similar to Perot. Another judgment distortion in a political context has been observed by Wood et al. [516] who collected data on 5000 candidates of the Greater London local elections held in 2010 (Figure 1.4) and showed that the order of candidate names on the ballot paper influenced the number of votes they received.

Several other distortions of human judgment exist. If we have recently heard news about a terrible plane crash, it may temporarily change our feelings about flight safety [132]; we mistakenly take a fictional event for a true memory [68]; a bogus personality feedback (e.g., astrology, fortune telling) often feels like an accurate tailored description [182]; after repeated exposure to a statement, the statement can be perceived as true (e.g., "People only use 10% of their minds." or "Eskimos have X words for snow."); as more and more people believe in an idea, others also "hop on the bandwagon"

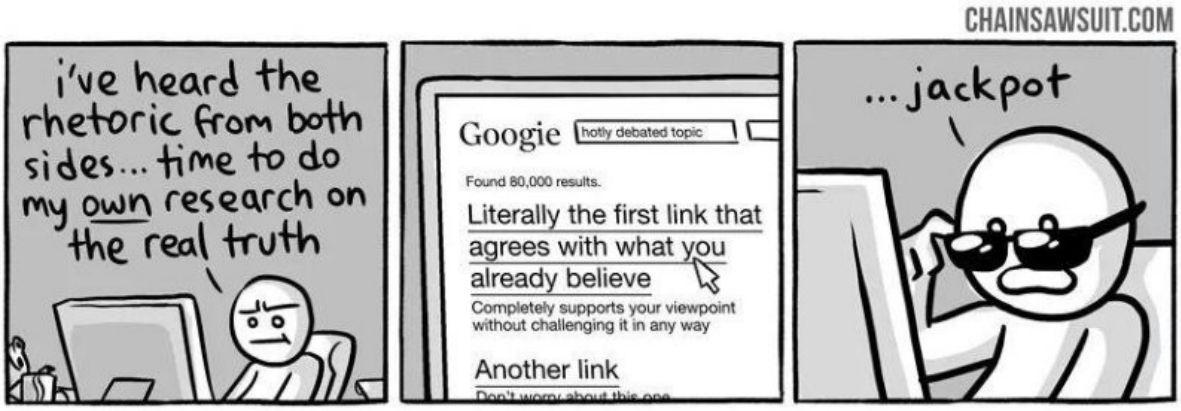


Figure 1.5: The confirmation bias. (source: chainsawsuit.com)

ignoring if true supportive evidence exists [350]; people are often influenced by stereotypes, expecting certain characteristics from a member of a group without prior information [314], and we tend to eat more food, if it is presented in a bigger container [184]. We often experience difficulty to understand risks and probabilities, e.g., how likely it is to win a gamble [474] or our team is to win a game [114], or the likelihood to have cancer after a positive mammography [41]. Moreover, once we have formed an opinion, we embrace information that confirms this opinion while rejecting, or simply ignoring information that casts doubt on it [352] (known as the "confirmation bias", see Figure 1.5). Our distorted views of reality inevitably affect our life choices: from the way we manage our finances to our actions on health prevention, to our political inclination, to the relationship with our fellows.

All these seemingly different distortions are known in the scientific literature under the same name: **cognitive biases**. A cognitive bias is not just a mistake, e.g., a miscalculation that people make due to a lack of education or understanding. Cognitive biases are erroneous behaviors that happen involuntarily [382], are quite robust even for intelligent and open-minded people, or even when they are proven to engage in effortful cognitive activities [506] or to have domain expertise [395]. Surprisingly, there are even cases where domain expertise showed to amplify cognitive biases [395]. Cognitive biases exist in many real-world contexts such as business, medical [117, 199], legal, or military settings [352] and most strategies that have been employed to mitigate them have appeared so far rather ineffective [30, 164, 262, 411].

It is often stated that people make poor and biased decisions because they have limited cognitive resources to process information, and they need to follow suboptimal strategies. Access to more information can indeed lead to better decisions, but it is not a bulletproof solution. In cognitive bias studies, participants often ignore the information given to them. Moreover, some biases, like the confirmation bias, demonstrate that people choose to filter the information in a way that will not challenge their prior beliefs. Even in our daily lives, Internet gives us information pluralism. Social and political diversity is often reflected in media content, various ideological opinions and viewpoints are discussed throughout a large range of information suppliers, such as channels, media companies,



Figure 1.6: Data visualization. source:dasaa.in

independent editorials, or blogs. To inform ourselves, we usually read newspapers, visit social media, listen to the news, read reports of experts, watch a video, or engage in discussions with our peers. Yet, the quality of our decisions is not always worthy of the volume of information available to us.

A common element among all these information sources is that the information is pre-processed, filtered and summarized by a third party. Therefore, it is on us to decide whether we can trust or be convinced by a source. As a consequence, we tend to believe and recycle the type of sources that do not challenge our current status quo. Otherwise, we feel lost on what information we can trust to help us make a well-informed decision. Perhaps the closest we can have to an original and "objective" information source, is to access the raw data that are often available online. Nowadays movements like the *open data* [511] advocate that all data should be freely available to everyone to use, while initiatives such as Data.gov gain more and more popularity. However, datasets of specialized domains can be very hard to understand, and often require training in statistics and comprehension of specialized measures. Even simpler datasets that contain cameras or cars are difficult to process in their original tabular formats. It is well known that there are ways of presenting/consuming such information that are less effective than others. A powerful tool that we can use to intuitively understand quantitative information without need for extensive training is data visualization [347, 495].

Visualizations can be powerful because they can effectively and faithfully convey accurate information. As many researchers emphasize, the main difference between visualizations and other disciplines that rely on visual representations such as art, videos in the news, infographics in a newspaper, or advertisements, is the focus on user efficiency [347]. In contrast, the emphasis of art is on conveying emotions, aesthetics or provoking thoughts; the emphasis of media is on telling a narrative; the emphasis of advertising is on selling a product [347]. Efficiency means to support the tasks of the users with respect to correctness, accuracy and truth [347]. Visualizations are not about "making pretty pictures" but they are meant to effectively assist users in their task [347]. In our case, this task is to make a well-informed and unbiased decision.

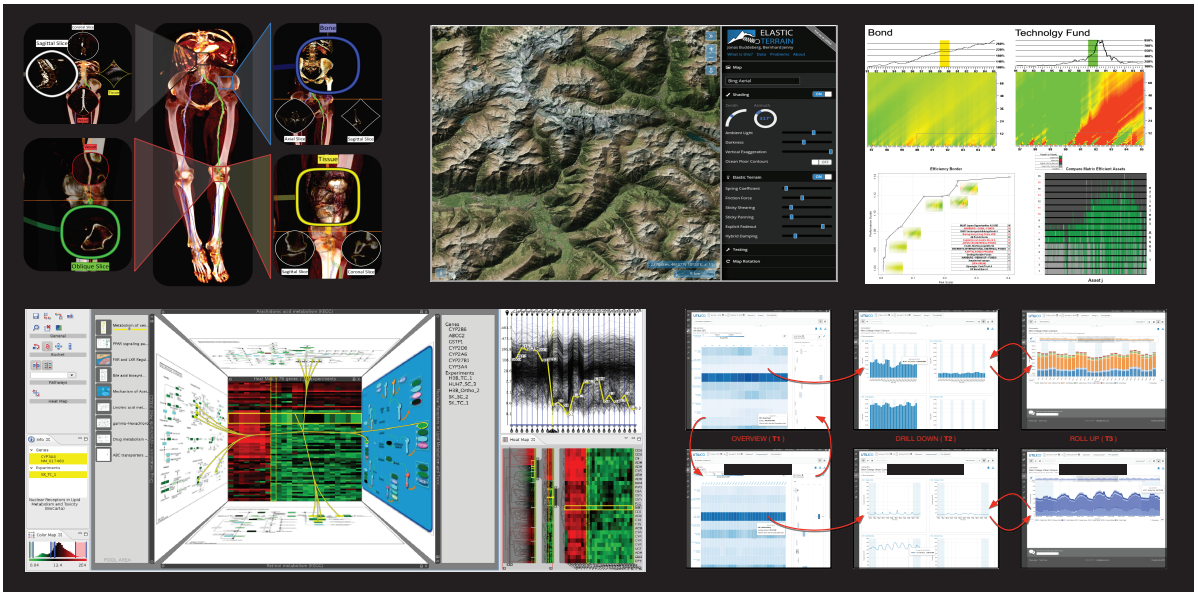


Figure 1.7: Examples of visualization systems displaying specialized datasets such as medical [337], geographical [77], financial [534], biology [302], and energy consumption datasets [71].

Unfortunately, there has been little work in visualization research on how to help individuals make decisions. Instead, visualization research mostly focuses on facilitating analytic activities. Visualizations are massively used by expert analysts who deal with complex datasets they need to understand, such as medical, geographical, financial, biology, and energy consumption datasets (Figure 1.7). There are also efforts in visualization research to make more "casual" [384] or "personal" [238] visualizations that can be understood by novice users (Figure 1.8). Casual and personal visualizations are meant to help users who can be characterized as *data enthusiasts*: nonprofessionals who want to analyze and get personal insights from a dataset. Ideally, what expert analysts and data enthusiasts have in common is their deep interest in the dataset itself. They want to dig into the data and gain insights. They are likely fascinated by rich data representations that can show them informative patterns and enhance their knowledge.

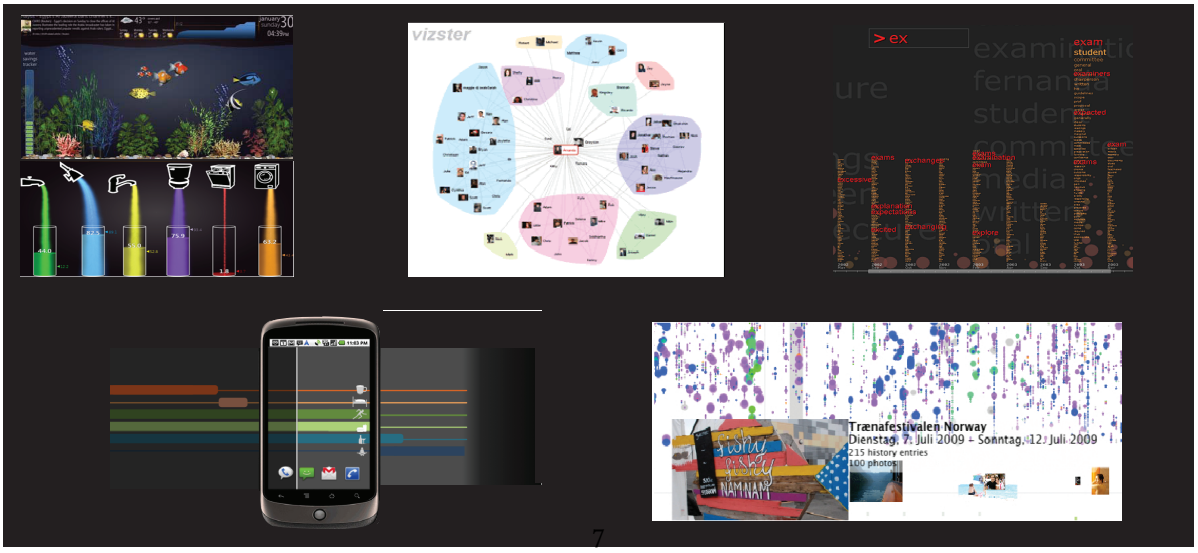


Figure 1.8: Examples of "personal" or "casual" visualization systems displaying personal datasets, e.g., water consumption [177], social networks[222], email [487], daily activities [48], music streams [49].

There are many reasons to think that current visualization tools, most of which were optimized for analytic activities, may not be optimal for helping a general audience make decisions. People who want to make well-informed decisions may not share a lot in common with analysts and data enthusiasts. One who searches a car that is both fast and affordable is not necessarily interested to know the range of car horsepowers or to observe the relationship between horsepower and acceleration. Similarly, when choosing an eco-friendly device, a representation that shows the distribution of energy consumption data may be unnecessary and too complex. Decision makers likely need a visualization tool to help them better manage information overload: filter out the unimportant, reorganize according to their preferences, and help them focus on the important. However, visualization tools provide limited interactions for such tasks. At the same time, a dataset may not provide all the information necessary for a decision. For example, to decide which politic strategy better represents one's own views, a dataset of economy indicators can be a good start, but likely needs to be combined with external knowledge that the decision maker already possesses or can collect. Most visualization systems do not support the externalization of this knowledge on the visualization itself.

Overall, information pluralism may not be a panacea, but the access to original sources with effective visualization tools can empower human decisions. Unfortunately, information visualization is not ready to fully address the challenges of visualization-supported decision making. Current visualization systems are not evaluated according to whether they can support decision making tasks. Moreover, there is little to no empirical evidence on whether visualizations are vulnerable to cognitive biases, or, instead, which are the visualization designs that could alleviate them. The goal of this dissertation is to widen the scope of information visualization in order to make visualizations accessible to all people who wish to make data-informed and unbiased decisions.

1.1 Thesis Statement

This dissertation makes a case for the following statement:

> ◇◇◇ To effectively support decision making, information visualization should move beyond the visual analysis paradigm. ◇◇◇

A **paradigm** [1] can be characterized as "*a framework containing the basic assumptions, ways of thinking, and methodology that are commonly accepted by members of a scientific community*" [5]. In this discussion, the **visual analysis paradigm** refers to the commonly accepted assumption in information visualization that the key purpose of visualization is to facilitate the understanding of information, such as the case of using a visualization tool to understand a complex dataset.

[1] Every word shown in this format: <**term**> indicates the definition of the term that will be used as such throughout the thesis. If next to the term there is a citation, the term is defined as such by the source (e.g., "**paradigm**"). If no citation is given, then the term is defined as such by the author of the thesis (e.g., "**visual analysis paradigm**").

Visualization research that targets decision support should shift the focus to the use case of using visualization to make a (good) decision. Moving beyond the visual analysis paradigm to support decision making would require thinking differently about a number of key concepts, including:

- ✓ Users — Visualization systems are often designed to assist professional analysts of any domain to analyze data and produce insights. "Casual" [384] or "personal" [238] visualizations have also been proposed to make visualizations accessible to novices. The latter work targets users who can be characterized as *data enthusiasts*: nonprofessionals who want to analyze and get personal insights from a dataset. This dissertation suggests to broaden the user profiles to *decision makers*: any user, decision making expert or novice, who may or *may not* be interested in understanding a dataset, but still needs visualization support to make data-informed decisions.
- ✓ Tasks — Information visualization is generally based on analytic tasks; either high-level tasks, such as confirmatory and exploratory analysis [274, 419] or low-level analytic tasks, such as value retrieval, identifying clusters and correlations [20, 504]. All these analytic tasks typically describe user activity aiming at data comprehension. Decision making is neither part of any visualization task taxonomy nor formally defined. Therefore, visualization task taxonomies need to be enriched with decision making tasks.
- ✓ Designs — Visual representations are usually designed with the criterion that the user understands the data presented. Nevertheless, decision making often involves subjective judgments and the way information is presented can also affect how decision makers evaluate their options. Visual representations need to be carefully designed not only as to be well understood, but so as to not distort a subjective interpretation.
- ✓ Interactions — Decision making is a rich process in which decision makers often follow various strategies and perform several iterations over the data, e.g., assign personal preference weights, reject unwanted solutions, evaluate their options, filter out information, add new information, compare, reweigh the options. Moreover, many decisions can be substantially based on external knowledge and subjective estimations, and, thus, some of these interactions may require direct modifications of the original data. Interaction needs to be rethought and visualization systems should enrich the interactions they provide to assist in all stages of the decision making process.

✓ Evaluation Currently, the evaluation of visualization systems is mostly based on whether users are able to use the system's interaction features, understand the data and finally gain insights. The evaluation of visualizations that target decision support should be primarily based on the outcome of the decision process. Visualization research lacks metrics and methodologies to assess decision quality, and it also lacks empirical evaluations of visualization systems for their ability to support decision tasks.

✓ Background Information visualization is a multidisciplinary field combining findings from fields such as computer graphics, visual design, and psychology. In particular the vision science has contributed a vast amount of perceptual principles and has inspired frameworks and design guidelines based on the limitations of human vision. Similarly, information visualization that targets decision support should also build upon findings from decision theory and behavioral economics involving models, decision strategies and limitations of human cognition.

1.2 Thesis Overview

Chapter 2 **Background** Chapter 2 reviews background work in decision theory and information visualization. First, Section 2.1 defines a decision making task and suggests its addition in the current visualization task taxonomies. The next two sections focus on how humans should, could and do make decisions. In particular, Section 2.2 covers the models which are used to formally describe the decision process, and Section 2.3 the systematic human limitations regarding this process, named cognitive biases. Section 2.3 provides a new taxonomy of cognitive biases based on the user task where such biases occur (the complete list is available in Appendix A). The last section 2.4 reviews the visualization systems that target decision-support, presenting their design and the methods used to evaluate their effectiveness in supporting decision tasks. The chapter finally discusses to what extent visualization research addresses the challenges presented in the decision theory sections.

Chapter 3 **Multidimensional visualizations for decision making** Chapter 3 attempts to address limitations in the evaluation of decision support visualizations as identified in Chapter 2. First, it conducts a systematic analysis to articulate the link between decision tasks and multidimensional visualizations, investigating the extent to which multidimensional visualizations are appropriate for decision tasks. Based on this analysis, it presents the first comparative evaluation of three multidimensional visualizations for their ability to support a decision: parallel coordinates, scatterplot matrices, and tabular visualizations. The chapter introduces decision metrics based on personal preferences, and a methodology on how to evaluate visualization techniques for decision support. It finally discusses the benefits and limitations of the methods and metrics used.

Chapter 4 **The role of instructions in decision making tasks** Chapter 4 investigates the role of instructions in decision making tasks and identifies differences in accuracy between decision and analytic tasks. In particular, this chapter explores the effects of providing task context in experiment instructions when evaluating visualization systems using crowdsourcing. The chapter examines whether a narrative component can engage and motivate users to give more accurate responses or if instead longer instructions induce more errors. Finally, the chapter discusses limitations in using objective dominance-based metrics to assess decision quality.

Chapter 5 **Detecting cognitive biases in visualization systems** : Chapter 5 presents the first study in information visualization research that detects a cognitive bias, named *attraction effect*, while using visualization systems. Reflecting on the findings about the limitations of human cognition presented in Chapter 2, this chapter considers metrics of decision quality that are different from the preference-based of Chapter 3 and the dominance-based of Chapter 4. In particular, this chapter investigates whether a decision can be "correct", yet irrational, in the sense that peoples choices can be influenced by irrelevant information. The chapter further discusses the implications of the attraction effect for information visualization designs.

Chapter 6 **Towards improving decision support visualization systems** Chapter 6 investigates how to improve visualization systems by helping people make better decisions. Section 6.1 suggests improvements based on empirical findings of previous chapters by presenting a novel decision support tool, named DcPairs. Section 6.2 classifies possible debiasing methods for decision support visualizations into educational, motivational, computation-aided, group-based and design-based. Section 6.3 follows a design-based approach, and investigates a novel debiasing interaction technique inspired by a well identified decision strategy. The section empirically verifies that such interaction technique can alleviate the attraction effect in DcPairs.

Chapter 7 **Conclusion** Chapter 7 summarizes the findings and contributions of this thesis and discusses perspectives for future research.

Chapter 2

Background

Scholarly books are often meant to provide a synthesis overview of a scientific field. One of the most well known, by Card, Mackinlay and Shneiderman in 1999, defines Information Visualization as *"the use of computer-supported, interactive, visual representations of abstract data to amplify cognition"* [84], and then goes on to list seminal research works that influenced the evolution of the visualization field (others focus on theory of human perception, e.g., Ware [495] or design principles, e.g., Few [162]).

One of the most recent influential books which, similarly to Card et al. [84], overviews seminal research works in visualization domain from the past 15 years is Munzner's "Visual Analysis & Design" in 2014 [347]. In the introduction, **Visualization** [347] is re-defined as follows: *"Computer-based visualization systems provide visual representations of datasets designed to help people carry our tasks more effectively. Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational* ***decision-making*** *methods."* The use of visualizations is considered essential in cases where human judgment is crucial to answer ill defined problems, when automated systems and statistics can not, all in *"an era that is characterized by the promise of better* ***decision making*** *through access to more data than ever before"* [347]. Therefore, decision making is a core part of data visualizations. This view is shared by many authors that consider decision support as a core challenge in visual analysis [274].

Munzner's book later summarizes among others user goals, evaluation methodologies, task taxonomies [347]. A careful reader may notice that the decision making references inside the book stop in the introduction chapter. Both in Munzner's [347] and also in the older visualization books [84, 162, 495], there is no explicit discussion or guidelines given on how to assess if a visualization system can assist users in making better decisions, e.g., what tasks or evaluation methodologies to use. For example, while visualization researchers have suggested taxonomies of user tasks (e.g., generate a

hypothesis, communicate, browse, [347] etc.), it is unclear which of these tasks should be considered when evaluating visualizations for decision support. Is there a task (or a combination of tasks) that is equivalent to, for example, "making a choice"? Alternatively, how can we evaluate the quality of a choice made with a visualization system, or which of the existing visualization systems can better support decisions?

There are three candidate explanations for why decision making is not discussed in scholarly visualization books more extensively:

1. too high level: Assessing the quality of a decision is indeed not explicitly addressed in the visualization field, because it is difficult to approach. Unlike a correlation task that has a definite correct answer, decisions are inherently subjective and concern ill-defined problems and trade-offs. Thus, it is hard to identify effective validation methodologies and metrics.
2. equivalent to visual analysis: Assessing the quality of a decision is addressed in the visualization field but under a different name (visual analysis and visual exploration). In other words, users who can perform visual analysis tasks (such as validate a hypothesis, identify trends, etc.) are also able to make good decisions. In a nutshell, visualization research considers data understanding as a sufficient condition for successful decision making.
3. selective content: Assessing the quality of a decision is indeed addressed in the visualization field, but, as in all overview books, the content is not exhaustive. Thus, the authors chose to not include existing work in evaluating visualizations for decision support.

This chapter will examine all three possible explanations. The first section attempts to operationalize decision making by defining a low-level decision task (section 2.1). Second, to examine whether successful visual analysis can be a sufficient predictor of a good decision, the following two sections will review background work in decision theory to better understand the inner workings of the human decision making process (section 2.2 and section 2.3). Third, to examine if indeed there is a lack of decision support methodologies, or if they are simply not covered in current texts, the fourth section will extensively review background work in visualizations targeting decision support (section 2.4).

2.1 Visualization Tasks

Visualization researchers emphasize that the main difference between visualizations and other disciplines that rely on visual representations (e.g., art, movies, info-graphics, advertisement) is the focus on effectiveness [347]. Effectiveness is determined as a corollary of the goal to support user tasks with respect to correctness, accuracy and truth [347]. On the other hand, the emphasis of art is on conveying emotions, aesthetics or provoking thoughts; of media on telling a narrative; of advertising on selling a product [347]. Visualizations are not about "making pretty pictures" but they are meant to assist the user effectively conduct her task [347]. In this dissertation, the goal of the user is to make an effective decision. Consequently, a concrete definition of what is a decision task and methodologies of how to support it are meant to be the backbone of this dissertation.

2.1.1 Analytic tasks

User tasks are usually collected through requirement analysis, or they are used as models during the design of the system, or they act like checklists during the evaluation of a system. Therefore, determining which are the core visualization tasks and forming task taxonomies is essential to visualization researchers.

Before defining a decision task, this section will first describe current common tasks in the information visualization field.

2.1.1.1 High-level analytic tasks

Interacting with visual data representations involves several high-level user tasks such as: (a) exploratory analysis, (b) confirmatory analysis, and (c) presentation [274, 419]. The exploratory and confirmatory analysis differ on whether the user has formed a-priori hypotheses when conducting data analysis. If not, the user conducts an exploratory analysis task, to search the data, analyze and finally identify useful information [274]. As soon as she forms one or more hypotheses, the user conducts a confirmatory analysis task, seeking to either confirm or reject these hypotheses [274]. Once the analysis is concluded, the aim is to communicate the result effectively in a presentation task [274].

To systematically determine how effectively a visualization system assists similar high-level tasks is rather challenging [274]. One way is to conduct a case study in a realistic set-up [379]. For example, focus on a particular application domain, such as weather forecasting, and ask expert weather-analysts to use a complex visualization system in order to determine what the weather will be for a given flight of interest [465]. Case studies – even though often criticized by some scientists [331] – can be very insightful since researchers observe users performing realistic tasks in situations that resemble everyday use of the system [379]. However, as Plaisant [379] reports, case studies can also be time-consuming, and their results are often not replicable or generalizable to other tasks or domains. Another way to determine the effectiveness of a visualization system is to conduct a usability study or a controlled experiment in which researchers often ask users to perform many low-level tasks [379]. Such low-level analytic tasks are presented next.

2.1.1.2 Low-level analytic tasks

Several taxonomies of low-level analytic tasks have been proposed [19, 20, 402, 445, 504]. One relatively recent and widely used is that by Amar, Eagan and Stasko in 2005 [20]. To define the tasks, Amar et al. used the term **data case** to refer to an entity in the data set, the term **attribute** to refer to a value measured for all data cases in a dataset, and the term **aggregation function** to refer to a function that creates a numeric representation of a set of data cases (e.g., average, sum, count).

Amar et al. [20] proposed the following low-level analytic tasks :

Task	Instructions	Example
Retrieve Value	What are the values of attributes X, Y, Z, ... in the data cases A, B, C, ... ?	What is the mileage per gallon of the Audi TT ?
Filter	Which data cases satisfy conditions A, B, C, ... ?	What Kellogg's cereals have high fiber ?
Compute Derived Value	What is the value of aggregation function F over a given set S of data cases ?	What is the average calorie content of Post cereals ?
Find Extremum	What are the top/bottom N data cases with respect to attribute A ?	What is the car with the highest MPG ?
Sort	What is the sorted order of a set S of data cases according to their value of attribute A ?	Order the cars by weight.
Determine Range	What is the range of values of attribute A in a set S of data cases?	What is the range of car horsepowers ?
Characterize Distribution	What is the distribution of values of attribute A in a set S of data cases ?	What is the distribution of carbohydrates in cereals ?
Find Anomalies	Which data cases in a set S of data cases have unexpected/exceptional values ?	Are there exceptions to the relationship between horsepower and acceleration ?
Cluster	Which data cases in a set S of data cases are similar in value for attributes X, Y, Z, ... ?	Are there groups of cereals w/ similar fat/calories/sugar ?
Correlate	What is the correlation between attributes X and Y over a given set S of data cases?	Is there a correlation between carbohydrates and fat ?

Low-level analytic tasks are meant to cover basic activities people do when analyzing data and have the following **limitations**:

- They do not necessarily *systematically cover* higher level tasks e.g., "Learning a domain" or "Predicting the future" [20];
- They are not necessarily *mutually exclusive*, e.g., to find an extremum value, a user may first sort the data cases [20], or to find an anomaly, the user may in some cases be looking for extremum values and in others for different patterns [20].
- They do not necessarily specify the *procedure* to complete the task (e.g., in order to find an extremum value a user may decide to sort all items or not).
- Some can be easily replaced by *computational methods* e.g., to find a correlation or a max value.

Low-level analytic tasks are not the reason why we need visualization systems. Visualizations are needed where there is room for human judgment, and automated computational methods cannot provide sufficient solutions [347]. The purpose of using low-level analytic tasks when evaluating a system is to act like a "checklist" [20] to assess its effectiveness. For example, if a user can effectively identify a correlation in a visualization system, it may be a stepping stone towards complex pattern recognition during exploratory analysis. Likewise, systems which fail to support low-level analytic tasks, are unlikely to assist high-level tasks.

2.1.2 Decision task

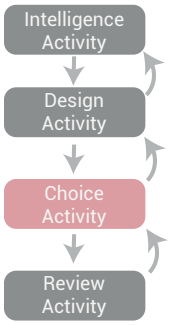
Similarly, a high-level decision task is presented. Later, the section defines a low-level decision task named "multi-attribute choice task". Multi-attribute choice tasks aim to enrich analytic task taxonomies so that they also cover the evaluation of visualizations targeting decision support.

2.1.2.1 High-level decision task

The elusive nature of decision making made it an apple of discord among several domains such as psychology, economics, cognitive science and management (e.g., arguing on whether decisions should be approached as an emotional process or as a set of actions driven by cost-benefit analysis). The goal in this dissertation is to operationalize decisions in an attempt to make decision tasks measurable, so that visualization researchers can understand their properties and the need to include them in their evaluations. This section borrows the definition of decision making from the management domain which focuses more on pragmatic decisions.

According to Simon (1960), the decision process is defined by the following four stages [432]:

Intelligence Activity	Search the environment for conditions calling for decisions
Design Activity	Invent, develop and analyze possible courses of action
Choice Activity	Select a particular course of action from these alternatives
Review Activity	Assess past choices



Primarily, visualizations for decision support attempt to aid the "intelligence activity" stage where people collect useful information and analyze the data of their environment. However, the decision making process does not stop at this information foraging stage, but it targets the selection of a single course of action, which it is often a far-from-trivial operation. Therefore, visual support should also be helpful in all stages, including choice activity and review of past choices.

A decision task as defined by Simon [432] differs from high-level analytic tasks, e.g., exploratory analysis, in that it serves different user goals: instead of finding useful information or confirming a hypothesis, now the goal is to choose a single course of action. However, both tasks involve ill-defined problems and complex analytic reasoning. Therefore, deciding on how to evaluate a visualization for its ability to support high-level decision tasks can be beyond doubt challenging. One way of going about it is, again, for visualization researchers to collaborate closely with real decision makers, e.g., conducting a case study in a business environment [427], and later validate the effectiveness of decisions taken in a real context. Nevertheless, time and lack of generalization remain important constraints. In this dissertation, the objective is to provide a feasible solution for more rapid evaluation of decision support.

As a figure of speech, to assess the decision equivalent of a correlation task for visual exploration, a low-level decision task is described next.

2.1.2.2 Multi-attribute choice task

The core of any decision task is described in the third stage of Simon's definition: choice activity, a selection of a particular course of action from a number of alternative actions [432]. Even though many types of choice activities exist, in the context of information visualization they are neither part of any task taxonomy nor formally defined. This section will provide an operational definition that is relevant to information visualization, and common in the decision making field.

In this dissertation, we [137] refer to **multi-attribute choice task** as a task that consists of *finding the best alternative among a finite set of alternatives, where alternatives are defined across several attributes*. One example is buying a camera at an online store, where each available camera is defined by its price and a number of technical features, e.g., size, weight, or resolution.

There is no unique way of defining a "good" alternative, and the best choice depends on the context. Nevertheless, "goodness" of choice can be defined in objective terms. For example, one can examine whether a decision maker follows certain principles of a normative model, e.g., dominance principle [477] (subsection 2.2.1 and chapter 4), or whether a decision maker is affected by factors irrelevant to the decision (chapter 5). "Goodness" can be approached also according to whether the decision maker followed a well-identified decision strategy (e.g., weighted additive strategy described in subsection 2.2.2). "Goodness" is often also defined in subjective terms, e.g., personal satisfaction with the choice or personal preferences (chapter 3). For now, it is noted that multi-attribute choice tasks can also involve subjective preferences and may or may not have an obvious "right" answer. Different possible metrics of goodness will be discussed in the next chapters.

Relevant concepts The notion of multi-attribute choice task is similar to the *preferential choice* previously mentioned by Bautista and Carenini in Information Visualization [51], as "the process of selecting the best option out of a set of alternatives". However, it is unclear of how the term "preferential" is used in this case. One possible explanation is that preferential refers to choices for which there is no external criterion to define which response is correct, and choice "goodness" depends solely on the decision maker's subjective goals [146]. It is further unclear how preferences are meant to be elicited. One way to elicit preferences is to consider them as "stated preferences" with questionnaires about hypothetical future choices. Another way is to consider them as "revealed preferences" through observation, where it is a-posteriori assumed that choices participants made reveal their underlying preferences [345]. In any case, the term preferential choice, as used by Bautista and Carenini [51], seems to involve only cases where choice goodness is defined by subjective preferences (e.g., choose a movie). Whereas the definition of multi-attribute choice task involves also cases where choice goodness is defined in objective terms (e.g., a data-driven medical decision that may involve external validation of some criteria).

The multi-attribute choice task term defined above is more closely related to the terminology of Multi-Criteria Decision Making (MCDM), a discipline that studies procedures to aid decision making in areas like business intelligence and finance [466, 535] (discussed in more detail in subsection 2.2.2). MCDM problems, however, refer to a broader class of decision-making tasks. Some MCDM tasks involve an infinite number of alternatives [466], and even when the alternatives are finite (referred to as Multi-Attribute Decision Making or MADM [466]), they are not necessarily known in advance [535]. Some MADM tasks can also involve ordering or classifying alternatives rather than identifying the best [466]. In this dissertation, the focus is solely on the task of finding the best among a finite number of alternatives, known ahead of time.

A key difference between the work in this dissertation and MCDM as a discipline is the focus on supporting spontaneous decision making aided by visualizations, *without any imposed procedure or strategy*. While MCDM methods are extremely useful for critical team decisions such as choosing a long-range business investment scheme [535], this dissertation focuses on how common visualizations can benefit a broad range of users without prior training in decision analysis. Thus it treats visualizations not as tools to guide users in their decisions, but rather as tools to help them better understand the information on which they base their decisions.

Multi-attribute choice task characteristics Since in a multi-attribute choice task all alternatives are *i)* known in advance, and *ii)* defined across a set of attributes, all information can be provided as a data table [361] where rows are alternatives and columns are attributes. Rows are also commonly called "data cases" in low-level analytic tasks (for example in Amar et al.'s taxonomy [20]), while columns are often called "dimensions" or "attributes". To help users understand this type of dataset, information visualization researchers have proposed a wide range of multidimensional visualization tools such as scatterplot matrices or parallel coordinates. Though many of these tools are used to analyze big datasets, most of them are also well adapted to the small datasets typical of common multi-attribute choice tasks (e.g., booking a hotel). Multi-attribute choice tasks and multidimensional visualizations are examined in detail in Chapter 3.

A multi-attribute choice task has the same limitations as with the low-level analytic tasks discussed in section 2.1.1.2. First, as in low-level analytic tasks, multi-attribute choice tasks do not necessarily *systematically cover* high-level decision tasks. Second, as in low-level analytic tasks, multi-attribute choice tasks are not necessarily *mutually exclusive* with other low-level analytic tasks. For example, to choose an alternative, one may need to derive values, determine ranges, identify an outlier (e.g., a cheap choice), check the correlation between price and quality. Third, as in low-level analytic tasks, a multi-attribute choice task does not necessarily specify the *procedure* to complete the task (e.g., one may choose the first alternative that satisfies her needs or review extensively all options). Finally, it can be again possible to replace some multi-attribute choice tasks with computational methods (e.g., computationally identify a single alternative that is superior to all others). Yet, a multi-attribute choice task differs from low-level analytic tasks in that it serves different user goals. The goal here is not

to compare values, sort, determine ranges or correlations; the goal is to select the single best among several possible alternatives.

2.1.3 Summary

This section suggested the addition of a low-level decision task named "multi-attribute choice task" in the current visualization task taxonomies. The following chapters may also use the term **choice task** to refer to a multi-attribute choice task for which the number of attributes is relatively small (one or two) or when there is no need to emphasize the number of the attributes (unimportant). The term **decision task** will be used to refer to high-level decision tasks (which also include choice tasks), or for decision tasks that do not precisely fit the definition of multi-attribute choice task as defined here (e.g., when it is possible to choose more than one alternative). The next two sections focus on how humans should, could and do make decisions. The first section mostly covers the models which were used to formally describe the decision process, and the second the human limitations regarding this process.

2.2 Decision Theory

People are faced with decision making every day. Some decisions can be mundane, such as what flavor ice-cream to choose, others more critical such as what cancer treatment a patient should follow or whether a person should choose to follow a political career or not. Decision theory is the area that studies the reasoning that underlies people's choices.

As we will see in the following sections, decision theory is divided into two branches: normative decision theory, which conceptualizes how people can make optimal choices given a set of constraints and values; and descriptive decision theory, which attempts to analyze how people actually make decisions.

2.2.1 Normative decision models

Human decisions are considered to be a task that is mostly subjective, context-based, and liable to unpredictable variations. However, early models of decision making did not approach it as such. These models assumed that decision makers have a fixed set of preferences, and make decisions in order to maximize their self-benefit by following certain principles of rational behavior.

One of the most well known rational models of decision theory is called "expected utility theory" by John Von Neumann and Oskar Morgenstern in 1947 [488]. Expected utility theory and its extensions [112, 168, 270, 371, 414] are based on certain rational principles. A subset of these principles is presented below [380] :

Completeness	Decision makers have well-defined preferences. They either prefer alternative A to B, prefer B to A, or are indifferent.
Dominance	Decision makers never select dominated alternatives. An alternative A dominates B if it is strictly superior in one attribute and superior or equal in all others. An alternative is dominated within a set of alternatives if there is at least one alternative that dominates it.
Cancellation	Decision makers ignore the identical attributes of alternatives. The choice is based on the attributes that differ.
Transitivity	If a decision maker prefers alternative A to B, and B to C, then that person prefers A to C.
Invariance	Decision makers are not affected by the way alternatives are presented.

Von Neumann and Morgenstern provided a mathematical proof that when decision makers violate principles such the ones described above, the expected utility [1] is not maximized [380]. Some of the most well-known extensions of expected utility theory were: the "subjective utility theory", which allowed people to make personal probability estimations [414]; and the "stochastic models of choice", which considered people's preferences as varying over time with random fluctuations rather than fixed to justify why people can be rational when they prefer alternative A one day and B the next [144].

According to most normative models, decision makers ought to explicitly calculate advantages and disadvantages of each alternative, including precise probability estimations. Reality though differs since information is often missing or uncertain, and people are not necessarily processing it in an expected way. For instance, when dealing with uncertainty, e.g., when choosing between an alternative with a certain outcome and a gamble, researchers found that decision makers tend to violate rational principles such as cancellation [17, 155], transitivity[472] and invariance [305]. Moreover, Herbert Simon in 1956 argued that decision makers do not necessarily make exhaustive comparisons of all alternatives to find the optimal, but rather choose the one that satisfies their most important needs [431].

2.2.2 Descriptive decision models

Since normative models do not adequately describe how people make decisions, many alternatives have been proposed known as descriptive models. Descriptive models attempt to describe "how people decide", as opposed to the normative ones which describe "how people should decide".

One of the most well known descriptive models is "prospect theory" developed by Daniel Kahneman and Amos Tversky in 1979 [265], which attempted to describe how decision makers behave with probabilistic alternatives that involve risk. Prospect theory states that decision makers are risk averse and also tend to evaluate alternatives according to a reference point rather than the alternative's true

[1] Utility: a measure of preference over some set of goods. Expected utility: a predicted utility value for one of several options, calculated as the sum of the utility of every possible outcome each multiplied by the probability of its occurrence

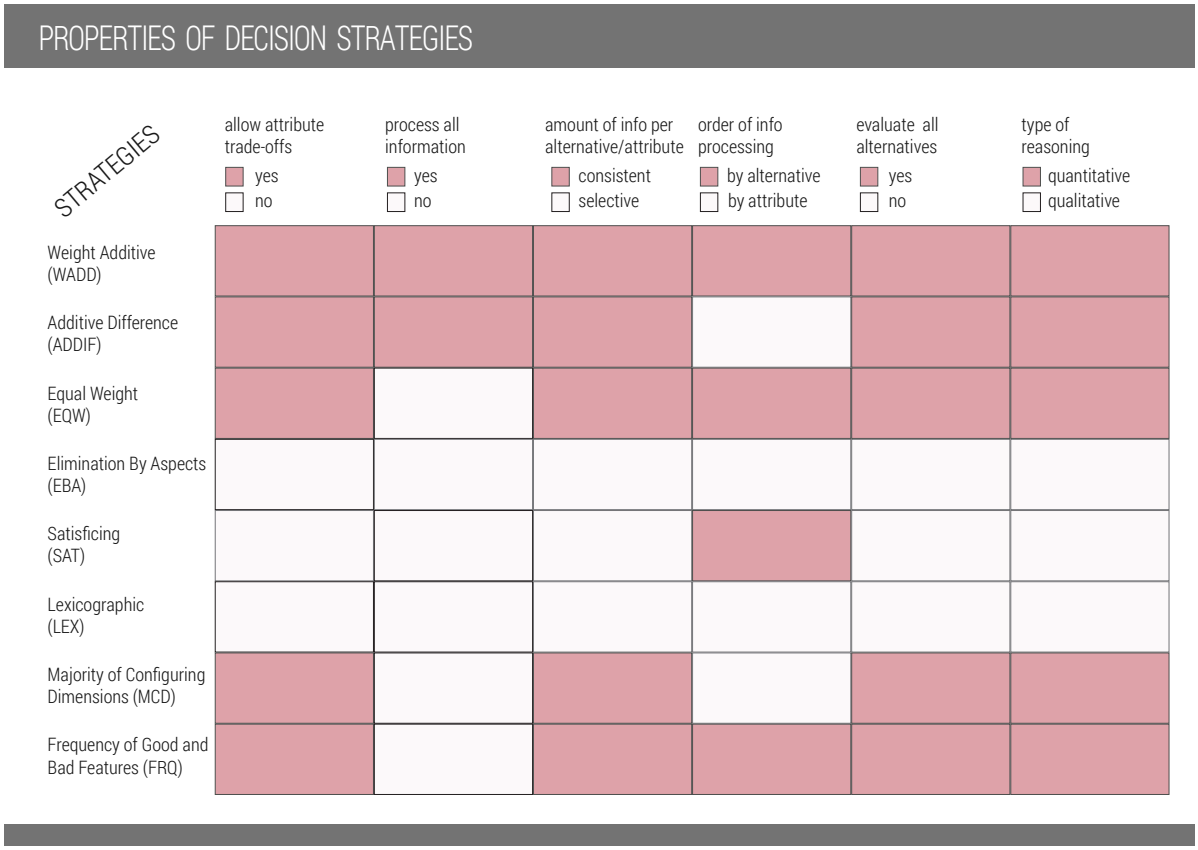


Figure 2.1: Descriptive models: properties of decision strategies (adapted image from Payne et al 1993 [372])

value. In particular, there is an asymmetry between losses and gains: e.g. decision makers tend to feel more the impact of a greater loss of $500, from a gain of $500, even though both involve the same amount of money. An alternative to prospect theory, "Regret theory", developed by David Bell [54] and Graham Loomes and Robert Sugden in 1982 [310], explains risk aversion by people's tendency to want to avoid the feeling of disappointment from an imaginary unfortunate outcome.

Nevertheless, choice goodness often depends on the type of task. Previous models only deal with outcomes defined along a single attribute (e.g., price) and a probability of the outcome occurring, and do not account for the fact that alternatives can have multiple conflicting attributes or do not need to have a probability associated with them. As previously explained multi-attribute choice tasks do not necessarily include a single best alternative. When dealing with conflicting attributes to optimize e.g. a negative correlation between price and quality, decision makers are faced with trade-offs and need to make compromises [151]. In these cases, it is no longer possible to consider a choice good if it is optimal (since objective optimal does not exist). Researchers have considered alternative criteria of goodness such as the consistency with people's goals [151]. Multi-attribute choice models describe

choice strategies to support such consistency depending on the problem type (list of decision strategies are seen in Figure 2.1).

In a task with two multi-attribute alternatives, decision makers often use "compensatory" strategies [380]. In compensatory strategies the superior value of one attribute compensates for the inferior value of another attribute [380]. One example of compensatory strategy is the "weighted additive (WADD)", in which the decision maker weights all attributes by their importance and chooses the alternative with the highest weighted value sum [372]. A simplification of the WADD is the "equal weight' (EQW)' strategy which sums alternatives without considering relative weights of importance [372]. Another strategy similar to WADD is the "additive difference (ADDIF)", where the decision maker evaluates first each attribute across alternatives, and weights and sums only their differences [380]. The "Majority of Configuring Dimensions (MCD)" strategy simplifies the ADDIF; the decision maker ignores attribute weights and codes a binary difference, considering only the direction of the difference and not its magnitude [372]. Finally, in the "Frequency of good and bad features (FRQ)" strategy, the decision maker assigns value thresholds to specify good and bad attribute values and then counts how many of these each alternative has. Depending on whether the focus is on good, bad or both the result can differ [372].

There are also non-compensatory models which do not allow trade-offs [380]; decision makers can drop an alternative with a bad value for one attribute, even if it has perfect values for the other attributes [517]. One of the oldest non-compensatory strategies is the "satisficing (SAT)" [431], in which the decision maker first, assigns attribute value thresholds, evaluates the alternatives by order of appearance, and chooses the first alternative that satisfies these thresholds. If no such alternative exists, the decision maker relaxes the thresholds and repeats the process, or chooses a random alternative [372]. Another non-compensatory strategy is the "lexicographic (LEX)" one, where the decision maker identifies the most important attribute, and then chooses the alternative with the best value for this attribute. If more than one alternatives exist, the decision maker identifies the second most important attribute and repeats[380]. Finally, the "elimination by aspects (EBA)" is a strategy proposed by Tversky in 1972 [471] that is very similar to the LEX, except that instead of choosing the best alternative for the most important attribute, the decision maker rejects all alternatives that do not satisfy a given threshold and repeats until only one alternative is left [380].

In order to better understand how these strategies compare to each other, a summary of their properties is presented in Figure 2.1. The difference between the previously discussed compensatory and non-compensatory strategies (all strategies in 1st column) is related to whether a decision maker has to deal with trade-offs, which can be emotionally uncomfortable [372]. Decision strategies differ in the information being processed. The most demanding strategies, WADD and ADDIF, force the decision maker to explicitly evaluate all alternatives (5th column) and to process all information relevant to the choice (2nd column). Other strategies reduce some of the information processed [372]. Moreover, decision strategies can differ on whether the decision maker has to process an equal amount of information for each attribute/alternative or not (3rd column). The decision maker can evaluate

alternatives across or within attributes (4th column) [372]. Based on this it has been suggested that attribute-based strategies are cognitively easier [406]. Finally, strategies have a different degree of quantitative and qualitative reasoning (6th column). For example, EQW requires summing of values, FRQ counts, and WADD multiplications. In contrast, strategies such as EBA are more qualitative in nature, involving only simple value comparisons [372]. Since less demanding strategies may not take into account information that is important to the decision maker, a common approach is to combine them. For example, a decision maker can use EBA to eliminate alternatives and then use WADD to choose between two or three [372].

2.2.3 Summary

Even though economists in the past tried to use normative models to explain consumer behavior, later studies showed that people rarely behave as a "rational economic man" [380]. Descriptive models have been proposed to better describe people's decisions, but still, most of these strategies require a lot of cognitive effort [372]. Especially under time pressure, people tend to switch to less optimal methods [528]. Besides, most of these strategies, – even though they were initially inspired by people's intuitive decision strategies – became rather formalized and often require lots of training [528].

As the reader may expect, the ambiguous nature of decision making is rather hard for any model to adequately describe. In an attempt to better understand the inner workings of people's decision behavior, the next section switches from a model-based to a task-based approach. In the task-based approach, people are usually given a small problem which, regardless of prior training or expertise, they tend to systematically fail to solve.

2.3 Cognitive Biases

As Munzner mentions: *"Visualization designers must take into account three very different kinds of limitations: those of computers, of humans, and of displays."* [347]. The following section is solely dedicated to the second. When making intuitive decisions people make approximations and employ unconscious heuristic strategies or rules of thumb. The imperfections of heuristics people routinely use manifest as cognitive biases [263]. Since visualization tools are used to support human decisions, the innate cognitive biases of people could also manifest in these tools. In order to understand how visualizations can support decision making, we need to understand the limitations in human reasoning. This section will present a detailed review of research in cognitive biases.

2.3.1 What is a cognitive bias

As discussed in the previous section, certain violations of normative rules, such as the ones of expected utility theory, could lead to irrational decisions. Those violations are named by the Nobelist Kahneman and by Tversky as "cognitive biases" [263]. However, nowadays the notion of cognitive bias is broader than what normative decision theories suggest, and includes deviations from various

types of norms, e.g., social behavior or memory retrieval. In particular, Pohl [382] provides a broader definition of cognitive biases distinguishing them also from other erroneous cases such as "typical errors", "misunderstandings" or, simply, an inability to recall a past event.

According to Pohl [382] a **cognitive bias** is a cognitive phenomenon which

1. reliably deviates from "reality",
2. deviates in a systematic fashion,
3. happens involuntarily,
4. is hard, if not impossible to avoid, and
5. appears rather distinct from the normal course of information processing.



Figure 2.2: Optical illusion.[11]

Regarding deviation from reality, in the case of a perceptual bias (as opposed to a cognitive bias), subjective perception would be compared to a concrete external stimulus [520] (Figure 2.2). In a cognitive bias, it is more difficult to define what constitutes a deviation from an objectively correct judgment or decision [382]. In fact, as novel theories in statistics or finance are emerging, some cognitive biases may disappear as noted by Gigerenzer [189]. More examples of deviations from reality are given in the next sections.

According to the second property of the cognitive bias definition, to characterize a simple error as a cognitive bias, it needs to deviate from a norm systematically and not at random [382]. To verify a systematic deviation, most cognitive bias experiment designs have a control group in which deviations are considered random errors, and an experimental group which should show a systematic effect [382]. If the bias requires repeated measures to detect, it is important to account for regression effects, that could lead to false result interpretations [382]. There are also cases where the bias is not necessarily observed in a single trial but appears only if the data are aggregated across a large number of participants and trials [382].

Another property of cognitive biases is that they happen involuntarily, without prior instructions or deliberate will [382]. People who are subject to a cognitive bias usually do not realize it is happening and they tend to believe that their decisions, judgments, and recalls are based on their available knowledge [382].

Consequently, the involuntary nature of cognitive biases makes them far from trivial to overcome [382]. They occur even when all relevant information is available and well perceived, and they often persist even when we inform or train people on how to overcome them [164, 199]. Bias alleviation techniques will be extensively discussed in chapter 6. For now, it is noted that debiasing methods are

known to be remarkably challenging.

The fifth property identified by Pohl is what probably makes cognitive biases appealing for researchers to explore. Cognitive biases appear to "stick out" as something that "piques our curiosity" [382] which is distinct from regular information processing practices. For example, they are beyond simply forgetting a piece of information, a commission error or a misunderstanding of a task [382].

2.3.2 Criticism of cognitive biases

The external validity of cognitive biases has been a subject of controversy in the research community. It is often argued that people may not be truly irrational because there is no information about the cost of people's errors compared to the cost of following normative principles [380]. Several authors, among them Gigerenzer and Brighton [192] argue that heuristics – the rules of thumb that people use and often result in cognitive biases – are very useful strategies in complex problems. People often achieve more accurate approximations when using a heuristic than by collecting an extensive amount of information and deploying complicated computations [192].

The heated debate on whether people are truly irrational has valid arguments from both sides, but it is outside the scope of this dissertation to address it. Research on visualizations, unlike disciplines of psychology or behavioral economics, does not focus on whether humans are good optimizers by nature. The objective of a visualization system is not to imitate the user's abilities but to empower them. Even if cognitive biases are the natural outcome of information overload, memory capacity and limited time, visualizations are known for aiding with these limitations. However, similar limitations, such as information overload, can also occur while using a visualization system, but these new extended boundaries need to be revisited by adding tool assistance to the equation.

2.3.3 Taxonomies of cognitive biases

One difficulty when transferring findings from decision theory to visualization is to diagnose which bias could be present when users perform tasks with visualizations.

Pohl [382] classifies cognitive biases into "memory", "judgment" and "thinking" biases. The *memory* class involves systematic errors in recalling or recognizing events [382]. The *thinking* class involves systematic errors in applying a certain rule (e.g., Bayes' theorem, hypothesis testing, syllogistic reasoning [2]) [382]. These rules come from several norms, e.g., probability theory, expected utility, or the falsification principle, which determine the actions that deviate from "reality". The *judgment* class involves systematic errors in subjectively rating a given stimulus (e.g., pleasantness, frequency or veracity) [382]. In judgment biases, people can often be affected by feelings of familiarity or confidence. As Pohl [382] also admits, this taxonomy has several limitations. Most biases in judgment and thinking

[2]Example of a false syllogistic reasoning: 1. " *All males are animals.*", 2. "*Some animals are aggressive.*" and 3. "*Some males are aggressive.*". The third conclusion here seems reasonable. Consider now replacing the second with: 2. "*Some animals are female.*". According to the previous rule, the conclusion should be 3. "*Some males are female.*" Thus, this is a wrong application of logic, because the animals which are aggressive are not necessarily men.

also involve memory processes such as encoding, storage, and retrieval [382]. Also, in the case where the material to memorize is outside of the laboratory, memory and subjective judgment biases cannot be distinguished because a faulty recall can be the reason for a faulty judgment (or not)[382]. Judgment and thinking classes also often overlap, e.g., people may not know that they are supposed to apply a Bayesian rule to estimate a probability and perform a subjective judgment of frequency.

Tversky and Kahneman classify biases according to which strategy (heuristic) we assume people follow to make a decision or judgment [474]. For example, some biases (most of them related to probability norms) are classified as outcomes of the "representative heuristic" where people estimate probabilities by the degree to which one event is similar to another event. For example, if they are given a salient description of an imaginary person named Linda with adjectives such as "bright", "outspoken", "deeply concerned with discrimination issues and social justice", and they are then asked to check off the most likely alternative "Linda is a bank teller" or "Linda is a bank teller and is active in the feminist movement", people tend to choose the second even though the conjunction of two events cannot be more likely (in terms of probabilities) than either event alone [380]. Another class includes the cognitive biases considered as outcomes of the "availability heuristic" in which people estimate an event as frequent or imaginable if they can recall it more easily in their minds and, neglect to applying a rational probability rule [474]. However, this strategy-based classification raised several criticisms by Gigerenzer [382], who considers these strategies as conceptually vague, imprecise and difficult to falsify, while other scientists give alternative explanations for why most of these biases occur [382].

Other classifications were developed in the domain of decision-support information systems. In 1986, Remus and Kottemann divided about 20 biases into two categories, data presentation and information processing, and later subdivided these categories based on the reasons why these biases occur (e.g., use of a certain heuristic, not understanding statistics, etc.) [393]. Similarly, Arnott in 2006, considered the nature of the cognitive bias and classified 37 cognitive biases into categories, examples of which are: *situation*, for biases related to how a person responds to the general decision situation or *confidence*, for biases that are believed to occur in order to increase the confidence of a person [31]. Arnott mapped each bias category with components of decision support system schema, e.g., data acquisition, processing, or output [31].

While they have not proposed a taxonomy of biases, visualization researchers have discussed important aspects of cognitive biases relevant to visual analysis. In particular, Zuk and Carpendale [536] categorize a subset of both biases and heuristics (without distinguishing them) based on how visualizations could mitigate them. For example, assuming that biases and heuristics happen because people have access to limited associations during the reasoning process (e.g., not available information requires memory retrieval), the authors argue that a visualization can give access to hundreds of associations in a given query and thus broaden the scope of the judgment. Ellis and Dix [154] also discuss biases that could affect visual analysis and emphasize the lack of work on investigating whether or not visualizations elicit cognitive biases in the viewer.

In general, the main limitation in all these classifications is that while biases are experimentally

verified, generic explanations of their nature, such as why the bias occurs or which heuristic people use, are mainly untestable and often topics of conflict among different scientists [382]. Moreover, the taxonomies based on information systems which associate biases with complex data processing are primarily based on a subjective interpretation from the author and not on sufficient empirical evidence.

For instance, most biases have been only verified in small puzzles of static textual representations and not while using a decision support computer system dealing with real datasets. Besides, most of these classifications include a small subset of biases 10-40, whereas in the larger (yet, not exhaustive) Wikipedia list there are 183 recorded cognitive biases up to now [510].

2.3.4 FAULTY: A new task-based taxonomy of cognitive biases

This section proposes a new cognitive bias taxonomy named FAULTY which is based on the *task* study participants have to perform, e.g., to solve a probability problem, to choose an item, to recall some material. The FAULTY taxonomy is not based on the explanation of why each bias occurs. For example, if study participants are given a task to estimate the likelihood of a heart attack or breast cancer, this will be considered a probability task. If study participants are asked to choose between different health insurances, this is considered as a choice task. Even though the origin of the bias can be important (e.g., false probability estimations can lead to false insurance choice), this explanation is not taken into account here. Also, the FAULTY taxonomy is not meant to group similar biases based on semantic interpretation – if two biases are known under different names and reported in different research works, these biases will be considered as distinct. However, if the task people performed was similar, the biases will be in the same category and therefore, easier for the reader to identify the similarities among some biases.

The procedure to develop FAULTY did not come from an existing cognitive bias taxonomy, but it was derived from an analysis process (e.g, similar to card sorting) to merge groups of tasks. This procedure was conducted as follows.

Step 1: Each bias was first searched based on whether it has been mentioned in InfoVis literature by typing the search term “bias name” + “information visualization” in Google Scholar. All InfoVis papers mentioning the bias (See Table A.1, column InfoVis) were collected. In the visualization papers mentioning a bias, the chosen reference was the source reference used to describe the bias and determined if it was an eligible source (see below). Only the first source was kept, in order to keep the total number of references manageable in this paper.

Step 2: In case no InfoVis paper mentioning the bias was found, or if these papers did not cite an eligible source, eligible sources were searched outside of the visualization literature, first in the Wikipedia list page, and then on the individual Wikipedia page of each bias. Again, the first eligible source was kept.

Step 3: In case no eligible source in Wikipedia was found, another source was searched by typing the search term “bias name” + “experiment" in Google Scholar. Only the first page of results was considered and examined the papers by decreasing order of citations. Again, the first eligible source

was kept. If no eligible paper was found, the Step 3 was repeated using a synonym for the cognitive bias. For each cognitive bias, the synonyms were collected on the individual Wikipedia page as well as in academic sources (see Table A.3) [3]

Step 4: In case Step 3 failed, we removed the bias from our list.

Source eligibility: A source was considered eligible if:

1. It was a peer-reviewed paper.
2. The document was accesible.
3. And the paper either:
   a) reported a human study that tested for the existence of the bias (full-text searches for the terms "experiment" and "study"),
   b) cited another paper that reports such a study, and described the paper's experimental task in detail.

Method b) was used when the original paper was too old for the document to be accessible, or when a peer-reviewed literature-review existed that described experimental tasks in enough detail. In general literature-review papers the favoured reference were the ones which provided a good overview of the different studies conducted on a particular cognitive bias. The accessibility rule (2) was applied only to help select one source over another, and no bias was eliminated because of that rule. The reliability of the experiment (e.g., experiment design, validity of statistical methods, size of the effect, etc.) was not examined.

Also, biases belonging to perceptual illusions (e.g., the contrast effect) were removed. Different cognitive biases pointing at the same sources were merged as synonyms.

[3] Steps 2 & 3 led us often to literature outside the InfoVis domain, which were kept as long as the eligibility criteria were met.

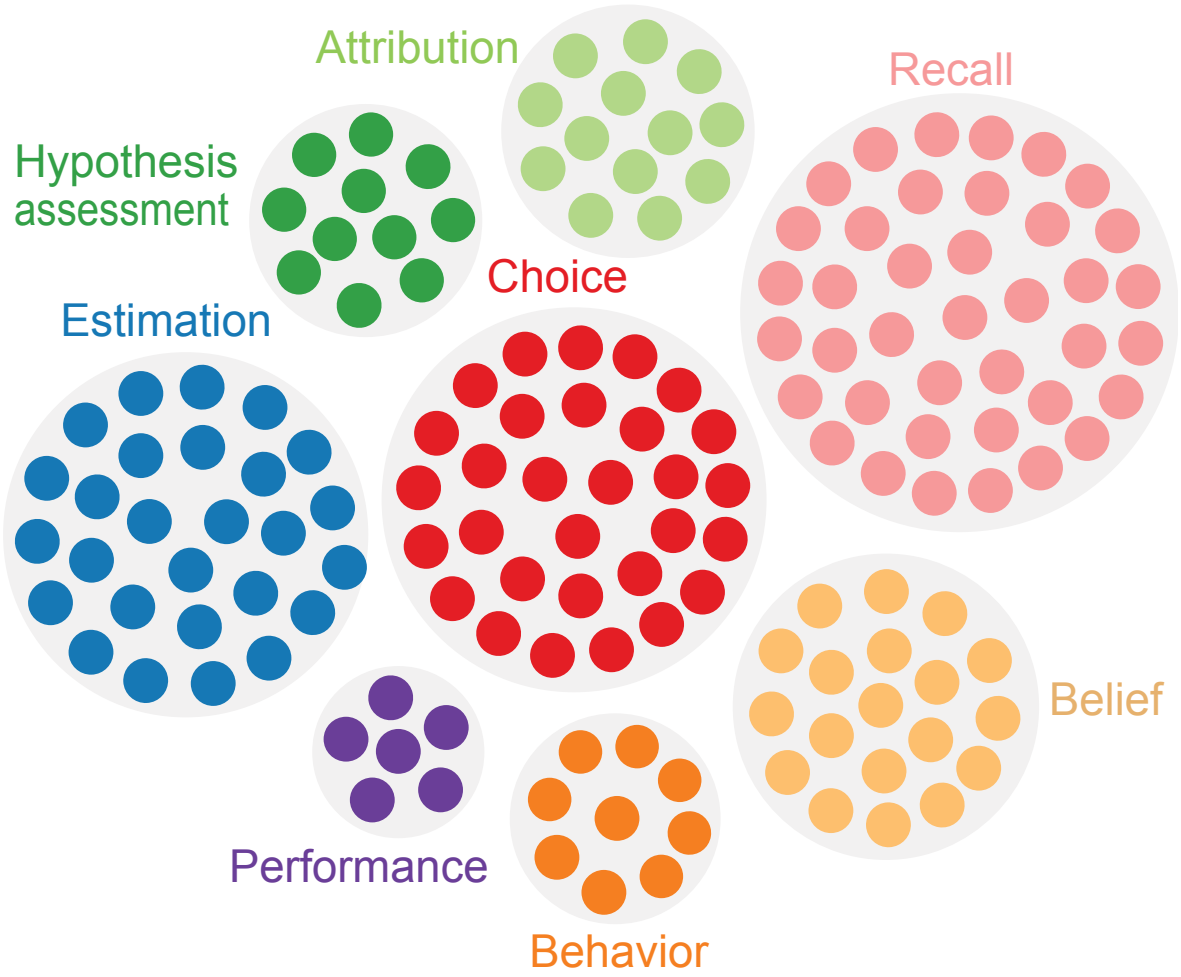


Figure 2.3: Overview of the FAULTY taxonomy. For the complete taxonomy, see Appendix A.

The FAULTY taxonomy (see overview in Figure 2.3) classifies cognitive biases in 8 categories, namely: **1)** *biases of a faulty choice task* (or simply, *faulty choice*) category, **2)** *biases of a faulty estimation task* (or simply, *faulty estimation*) category, **3)** *biases of a faulty recall task* (or simply, *faulty recall*) category, **4)** *biases of a faulty hypothesis assessment task* (or simply, *faulty hypothesis*) category, **5)** *biases of a faulty attribution task* (or simply, *faulty attribution*) category, **6)** *biases of a faulty performance evaluation task* (or simply, *faulty performance*) category, **7)** *biases of a faulty belief task* (or simply, *faulty belief*) category, and **8)** *biases of a faulty behavior* (or simply, *faulty behavior*) category. Each of these categories will be discussed in detail in the following sections. The sections will also describe a subset of cognitive biases. The complete table with the FAULTY taxonomy is available in Appendix A. Even though the following sections may provide some pointers to the Appendix A (e.g., color tags, hash-tags), alternating section and appendix is not necessary to follow the text. Instructions on how to read the Appendix (including color tags and hash-tags) are also given in Appendix A.

Unlike previous taxonomies that were based on untested, hard to grasp and often conflicted

explanations of why a cognitive bias occurs, this taxonomy is organized by task. Therefore, visualization designers can look at which biases may exist in their system, assuming they know the tasks users will perform, rather than trying to guess the inner cognitive process the users may follow. For example, the topic of interest in this dissertation is decision making. Thus, the *faulty choice* category reveals the biases which are likely to appear while users perform a multi-attribute choice task. Similarly, visualizations research targeting memorability of visualization designs [62, 215] may want to consider the biases of the *faulty memory* category; researchers who study confirmatory analysis tasks [274] could be more interested in the *faulty hypothesis* category, and researchers working on uncertainty visualization [536] may want to focus on the *faulty estimation* category. Moreover, the FAULTY taxonomy outlines how the biases were tested by giving pointers to the original experiments, which may help visualization researchers to replicate evaluation methodologies that are well-established from the psychology field. In addition, even though it is very likely that additional biases exist which are not included in the list, this taxonomy is, to the author's knowledge, by far the largest in the literature including biases from different domains (e.g., psychology, consumer research, sociology). Usually, taxonomies developed by researchers of one domain do not account for biases of other fields. Finally, a task-based classification of cognitive biases can potentially underline new patterns by presenting biases from a different angle. For example, similarities between some tasks may reveal biases with the same root.

2.3.5 Biases of a faulty choice task

The **biases of a faulty choice task** category (■ #CHOI in Appendix A) includes the systematic errors that occur when participants are given a certain *choice task*. This type of experiments are often referred to in psychology as *choice studies* in which participants are "*required to exhibit a preference for one of the several stimuli or make a different prescribed response to each of them*" [37]. A subset of these biases is discussed here and a list of 23 #CHOI biases is available in Appendix A.

Some faulty choice biases occur when people are dealing with decisions under uncertainty. For instance, people tend to avoid choices associated with ambiguous outcomes [176], and if the choice set contains any certain (even if not optimal) alternative they tend to stick to it [44]. Moreover, people often show different preferences for gains (e.g., allowances) or losses (e.g., prohibitions) [418] (known as *loss aversion*) even if these alternatives are only framed as such (*framing effect*). For example, in an experiment where people were asked to choose a program to combat an unusual Asian disease, the program framed as a "33% chance of saving a life" was preferred over the program of "66% chance of death", despite that the result would be the same [475].

Visualization researchers often discuss that choice biases under uncertainty can have important implications in visual analysis [128, 154, 405, 536]. However, as Ellis and Dix [154] point out, there is very limited empirical work to identify these implications. One example that could be considered a visualization from the HCI domain is the

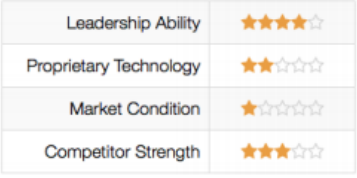


Figure 2.4: Zhang et al. [530]

work of Zhang et al. [530], which showed that startup companies
presented with tabular forms (Figure 2.4) tended to be subject to loss aversion bias.

Nevertheless, not all choice biases involve uncertain outcomes. When people have to choose one alternative over the other, they are often unconsciously influenced by factors irrelevant to their decision. In most biases, decision makers do not evaluate alternatives in isolation, but based on the context in which the alternatives occur [380]. One well-studied example of such a bias is the *attraction effect*, where one's choice between two alternatives is influenced by the presence of irrelevant (dominated) alternatives [240]. This bias will be discussed in detail in Chapter 5. Other context-based biases involve whether alternatives are presented separately or juxtaposed [237], or among more extreme [433], unavailable [376] or more familiar [527] alternatives. Other cases of biases of a faulty choice task involve attachment to alternatives for which people can receive an immediate reward [460], or for which they had previously invested manual effort [188, 355]. Examples also include attachment to alternatives which people owned in the past [282]. and people who avoid making any choice that will deviate from their current status quo [408].

An example of an interesting choice bias in government elections was identified using visual analytic techniques. Several scientific studies had long investigated the hypothesis that the order of candidates in the ballot papers can affect the result of the elections, but they only found inconclusive evidence. Wood et al. [516] collected data from 5000 candidates of the Greater London local elections held on the 6th May 2010, analyzed them using hierarchical spatially arranged visualizations, and showed that the position [4] of candidate names on the ballot paper (shown in Figure 2.5) indeed influenced the number of votes they received. Wood et al.'s work showed that an alphabetical tabular representation of candidates can contribute to biased election results.

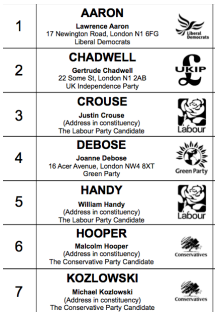


Figure 2.5: Wood et al. [516]

The *biases of a faulty choice task* category is particularly important for this dissertation, because it reveals the biases that are more likely to appear while users perform a multi-attribute choice task. According to this observation, a cognitive bias of this category (i.e. attraction effect) will be further investigated in Chapters 5 and 6.

As a final note, in almost all previous studies the cognitive bias was elicited by alternatives shown in textual formats, oral descriptions and static data tables. It is often stated that the reason behind biases is that people have limited cognitive resources to process information, and they follow suboptimal strategies. It is possible that decision support visualizations, which are known to enhance human cognitive abilities, could help overcome these limitations. However, this question is still largely underexplored.

[4] The ballot names bias shares some similarities with the *serial-positioning effect* [348] where people better recall the first (primacy) and last (recency) items in a list. However, it is not the same bias since the task in ballots is to choose (a candidate) and not to recall her name. It is possible, though, that people chose the candidates who were easier to remember.

2.3.6 Biases of a faulty estimation task

Cognitive biases observed in choice tasks are not the only ones which can lead to poor decision making. Decisions often require people to make accurate estimations, for example, to estimate the likelihood of a robbery (to choose whether to insure a car), or to estimate possible retirement needs (to choose a supplementary program or a health insurance). The **biases of a faulty estimation task** category (■ #PROB and ■ #SEST in Appendix A) includes the systematic errors that occur when participants are asked to make an estimation of a future event occurring (prediction) or an estimation of frequency. A list of 23 estimation biases is available in Appendix A.

Most estimation biases involve a probability problem in which participants are expected to apply a certain rule (e.g., Bayesian statistics) (■ #PROB) and make a calculation. For example, people tend to overestimate the likelihood of an event (e.g., having breast cancer after a positive mammography) because they compute probabilities about a specific case, ignoring overall probabilities that apply to the general population [41, 476], or they misinterpret conditional probability rules [56]. Moreover, people often do not revise their estimations effectively in the light of new information [147, 149**?**]. Researchers have studied how visualizations such as Euler diagrams and frequency grids (Figure 2.6) can reduce such biases in probabilistic reasoning [276, 332]. Even though these studies did not observe a systematic error, Micallef et al. [332] showed that visualizations might improve the overall understanding of probability problems.

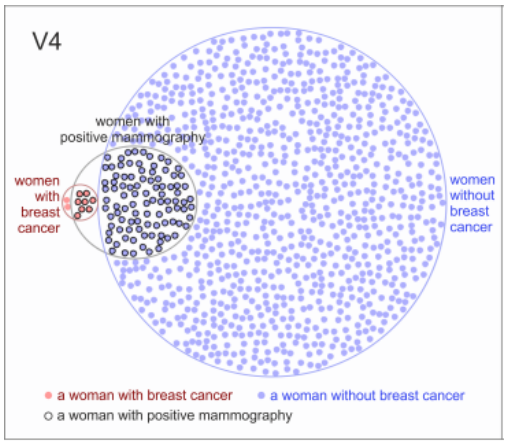


Figure 2.6: Micallef et al. [332]

Other estimation tasks involve *qualitative* estimations of probabilities (■ #SEST). The correctness of such tasks is not based on precise computations like before, but researchers usually examine factors that may irrationally influence an estimation. One interesting example is the *anchoring effect* [474], where the estimation is biased toward an initially presented value. That value can be irrelevant to the task. For example, Tversky and Kahneman asked people in an experiment to spin a fortune wheel, and, just after, asked them to estimate the number of African countries in the UN [474]. Their estimation tended to be a number close to the one that the fortune needle landed on [474]. People also tend to make more optimistic predictions for themselves than for others about future events (e.g., to find a dream job, not getting a divorce or lung cancer) [505], or about how much time they need to complete a future task [78], and they often treat serious harmful risks as nonexistent [452]. Other qualitative cases are mis-estimations of frequencies, e.g., estimating wrongly that words starting with the letter "R" are more frequent than words having the "R" in the third position. This is believed to occur because people make estimations using the "availability heuristic" in which the first type of word are easier to retrieve from the memory and, therefore, they are perceived as more frequent [473].

2.3.7 Biases of a faulty recall task

The **biases of a faulty recall task** category (indicated with ■ #MEMO in Appendix A) includes the systematic errors that occur when participants are asked to recall or recognize previous material. A list of 39 #MEMO biases is available in Appendix A.

Memories are not exact copies of past experiences stored in a warehouse. Instead, people construct their memories at the time of withdrawal [380]. For example, there are cases where post-event information influences the ability of people to accurately recall the event, known as *misinformation effect* [38]. Moreover, people tend to better recall visual representations over words [325], auditory information over visual information [197], self-generated content over read content [437], pleasant over unpleasant emotions [492], interrupted tasks over completed ones [152], humorous [451] or bizarre items [327], and information hard to comprehend [115] or hard to find through a search engine (known as the *Google effect*) [440]. People also tend to consider material retrieved from their memory as an original inspiration [75], which can be an unintentional cause of plagiarism. Conversely, people consider some imaginary events as real [68], a phenomenon often observed in crime witness interviews by misleading suggestions [306].

Even though the memorability of visualizations has often been associated with their effectiveness [62, 215], it seems that there are no previous works investigating memory biases in the context of visualizations. Memory limitations, though, are believed to be a fundamental source of several other cognitive biases [263] (e.g., some choice biases discussed in previous section).

2.3.8 Biases of a faulty hypothesis assessment task

The **biases of a faulty hypothesis assessment task** category (indicated with ■ #HYPO in Appendix A) includes the systematic errors that occur when participants are asked to investigate whether a hypothesis is true or false. The term hypothesis here does not necessarily refer to formal statistical hypothesis, but any statement, informal or formal, that can be either confirmed or disconfirmed (e.g., via reasoning, previous knowledge). A list of 12 #HYPO biases is available in Appendix A.

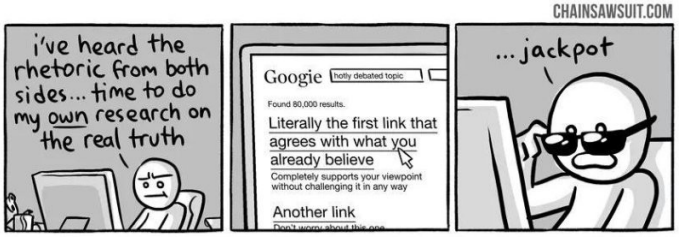


Figure 2.7: Confirmation Bias. source:chainsawsuit.com

When people investigate a hypothesis they tend to favor any strategy (e.g., the way they search for information, the way they reason to drive conclusions) that can confirm this hypothesis and subconsciously ignore any disconfirming evidence [352]. This is known as the "*confirmation bias*" and it is considered as one of the most impactful cognitive biases in human reasoning. As Nickerson puts it, "*If one were to attempt to identify a single problematic aspect of human reasoning that deserves attention above all others, the confirmation bias would have to be among the candidates for consideration. Many have written about this bias, and it appears to be sufficiently strong and pervasive that one is led to wonder whether the bias, by itself, might account for a significant fraction of the disputes,*

altercations, and misunderstandings that occur among individuals, groups, and nations."[352]. Other similar cognitive biases occur when a person falsely considers a hypothesis as true after repeated exposure to it [213] or without also testing indirect additional hypotheses [496], or when she considers a relationship between variables that does not exist [89].

Other cases of biased strategies in hypothesis assessment have been observed when a critical experiment variable is typically ignored by experimenters: the experimenter himself [398]. One of the most representative bias is the "observer-expectancy effect", where the experimenter subconsciously influences the participants to behave in the way that confirms her hypotheses [398]. A very first example of such involuntary cuing was studied around 1904 on the so called "Clever Hans," a horse claimed by his owner to be able to respond to human language and make mathematical calculations by tapping his hoof (e.g., asked 3 plus 2 and tapped 5 times). Oskar Pfungst discovered that the horse was, in fact, responding to subtle physical cues (e.g., postural adjustment, or slight head movement), which were nevertheless not intentional signaling since the horse was responding correctly even with other questioners besides the owner. Pfungst noted that when a questioner did not know the answer, the horse could not respond. The horse was indeed exceptionally clever, not due to his linguistic or math ability, but because it could perceive very subtle muscle movements [398]. Pfungst's study was the first to raise the question whether it would be possible for human participants to be similarly affected by the experimenter's hypotheses, given that a horse's behavior could be affected by the observer's expectations. [398].



Figure 2.8: Observer-expectancy effect. source:quibb.com

Most of these biases are often mentioned in the visualization literature as critical domain challenges [154, 536] but there has been no empirical verification of a systematic error during visual exploration or empirical evaluation of visualizations. Likewise, it would be interesting to investigate how visualization techniques could alleviate, for example, confirmation bias (e.g., by offering alternative representations).

2.3.9 Biases of a faulty attribution task

The **biases of a faulty attribution task** category (indicated with ■ #ATTR in Appendix A) includes the systematic errors that occur when participants are asked to provide explanations of their own or other people's behavior. Attribution is a term borrowed from the field of psychology. Biases in this category were also explicitly identified in the psychology literature as "attribution biases". Thus, this dissertation did not contribute to originally identify this category. A list of 13 #ATTR biases is available in Appendix A.

In most of these biases, people tend to favor themselves over others in the explanations they give. In order to explain why a joined achievement was successful, people tend to overestimate their

contribution [401] as the fair result in terms of ability and effort, and do not have a fair assessment of their peer's achievements [83]. When it comes to failures, they tend to attribute their own to situational factors, and the failures of others to personality weaknesses [260]. Similarly, they attribute their job motivation to the prospect of developing new skills, and the motivation of others to monetary rewards [216]. In order to evaluate a joined action, people are also more likely to assign themselves more varied personality traits, whereas their view of their peers is less varied, but rather uniform [387]. Finally, people often justify their judgments as being typical and publicly accepted [400].

Similar attribution biases occur when people have an unrealistic justification involving in-group and out-group actions. For example, when Hindu office clerks evaluated random, unknown profiles of Hindu and Muslim populations, tended to assign characteristics such as "hospitable", "kind" and "honest" to the former and "rude" or "cheater" to the later [377]. When attributing the outcome of joined out-groups action, they also tend to overgeneralize individual behaviors of their group [208]. And, conversely, they tend to infer decisions made by a group to individual people (e.g., action of a whole nation to an individual citizen) often ignoring conflicted information on the subject [18]. Moreover, people often tend to attribute harmful action to others (e.g., to consider them as responsible for negative gamble results), or to attribute their ambiguous behavior to intentionally negative reasons (e.g., I see my peers laugh, they may laugh with me) [138].

It is not always clear whether attribution biases are indeed inevitable or if they can be alleviated by activating background knowledge. Sometimes people in previous experiments tended to ignore the information available to them in textual formats (e.g., facts about a population or descriptive profiles of individual people). However, it would be interesting to investigate if the quality of the given information plays a role. For example, a newspaper article may not have the same effect as a user-driven visual data exploration.

2.3.10 Other categories

The **biases of a faulty performance evaluation task** category (indicated with ■ #PERF in Appendix A) includes the systematic errors that occur when participants are asked to rate their performance after a given problem. Most of these biases are related to overconfidence where participants rating was higher than their accuracy [280]. Confidence can change according to the difficulty of the task (overconfidence for hard tasks, conservatism for easy ones [304]) or the expertise of the participant (overconfidence in non-specialists, conservatism in experts [288]). User confidence is an important metric in information visualization, and it is often associated with the understanding of a visualization tool [142]. However, according to the findings of cognitive biases research, the reliability of confidence metrics needs to be investigated together with task accuracy. A list of 6 #PERF biases is available in Appendix A.

Another category is the **biases of a faulty belief task** (indicated with ■ #BELI in Appendix A) where participants make systematic errors when answering questions regarding their opinions on political, moral, social beliefs about certain situations or individual people. For example, people's

reported beliefs on issues such as abortion or sovereignty can change according to the majority opinion, known as "Bandwagon effect"[350] or people often assign moral blame depending on the outcome, not on the action [125]. For example, not wearing a seatbelt is more irresponsible, if an accident happens. Notably, people tend to tend to believe that other people are more biased [289] and more affected by mass media propaganda [25] than themselves. Moreover, people and also tend to report certain characteristics from a member of a group (e.g., race, ethnicity, gender, age) often ignoring any conflicting evidence, known as "stereotyping"[314]. Even though there may appear to be some similarities between the belief and the attribution category, they have an important difference: in the attribution category, people's task is to explain or reason about a phenomenon (e.g, USA has economic growth, because Americans are smart), whereas in the belief category, they report an opinion (e.g. I believe that Americans are generally smart.). Similarly, if people make a prediction based on a belief, this bias will belong to the estimation task category (e.g. USA will likely grow – because Americans are smart.). All these examples illustrate different cases, since people who have certain generic beliefs about certain topics will not necessarily reason or predict the future based on these beliefs. A list of 17 #BELI biases is available in Appendix A.

The last category is the **biases of a faulty behavior** (indicated with ■ #BEHA in Appendix A) where participants are not instructed to perform a specific task, but the study examines their behavior in certain situations. Examples of such behavior are people who eat more food in bigger containers [184]; group meetings in which members discuss known facts more extensively than new information [40]; investors who do not monitor their portfolios frequently enough when they show negative information [269]; or cases of a physical disaster where people do not evacuate [292]. There are certainly connections between categories. For instance, it is possible that people who do not evacuate in a dangerous situation (e.g., earthquake) failed to predict an event (e.g. the building is likely to collapse). However, since people are not asked to report a prediction, it is unclear whether the reason for inaction is a false probability estimation. A list of 6 #BEHA biases is available in Appendix A.

The last two categories are inherently linked to moral beliefs and human instinctive behavior, which lay beyond the scope of this dissertation. However, similarly to all other bias categories, the possible connection of such errors to visualization systems is an unexplored topic. For example, is it possible that user-driven data exploration of criminal records could alleviate a stereotyping bias? Similarly, should negative information be visualized in a way that will not escape the attention of an investor?

2.3.11 Conclusion

This section classified 139 cases where people systematically and involuntarily deviate from what is expected to be a rational "reality". For example, their choices are often influenced by reasons irrelevant to the objective qualities of the choice alternatives. Cognitive biases are often mentioned as important in the visualization literature [536]. Nevertheless, it seems that there is no visualization study that provides evidence for the alleviation of a cognitive bias. Moreover, it seems that there is only one study

[530]) that provides some evidence for the existence of cognitive biases in visualizations, but as it can be observed in Figure 2.4, the visualization design was rather rudimentary. There are some works that examined cognitive biases in the context of visualizations [67, 276, 332, 516], but they did not provide evidence for neither alleviation, nor detection of the bias while using a visualization tool.

As explained in Section 2.3.1, cognitive biases are a complex type of error in the sense that the deviation from "reality" needs to be observed in a systematic fashion, to happen involuntarily, must be very hard to avoid, and must appear rather distinct from the normal course of information processing. Nevertheless, simpler errors (resulting from, e.g., miscalculation or misunderstanding of information) are also important and common, and they have been more widely studied in visualization. The next section will examine such errors in visualizations that target decision-support. In particular, it will examine how visualization researchers evaluate their effectiveness in supporting multi-attribute choice tasks.

2.4 Decision Support Visualization Systems

This section covers two types of decision-support visualization systems: (i) visualization systems explicitly designed to support decisions (e.g. Value Charts [51]) or (ii) general purpose visual analysis systems that they have been illustrated using a decision making scenario (e.g., choose a camera use case in ScatterDice [156]). The complete list of the decision-support visualization systems is available in Appendix B.

2.4.1 Overview

The first subsection briefly reviews dataset-dependent systems and the second presents generic decision support visualization systems in more detail.

2.4.1.1 Dataset-specific visualization systems for decision support

This dissertation focuses on how to support a generic multi-attribute choice task in any dataset, which it is not fully supported by dataset-specific systems. Dataset-specific can be considered systems whose design is hard to generalize beyond their specific application domain (indicated with the white space in the "I"column in Appendix B). For example, FinVis [405] is a tool that shows investment options in a risk plot, along with the overall aggregated risk as a Gaussian gradient, to help financial decisions (Figure 2.9 a). This design is not straightforward to apply in a choice task that uses a standard InfoVis dataset, e.g., cars or cameras datasets [156].

A range of dataset-specific visualizations have been proposed to help people make decisions in several application domains. Stratos (Figure 2.9 b) [35] helps software project managers select which features to include in each development stage during software production, by simultaneously visualizing all possible software release plans. Other systems use visualizations to communicate financial risk helping users find profitable investments, such as FinVis [405], Financial Portfolio [415],

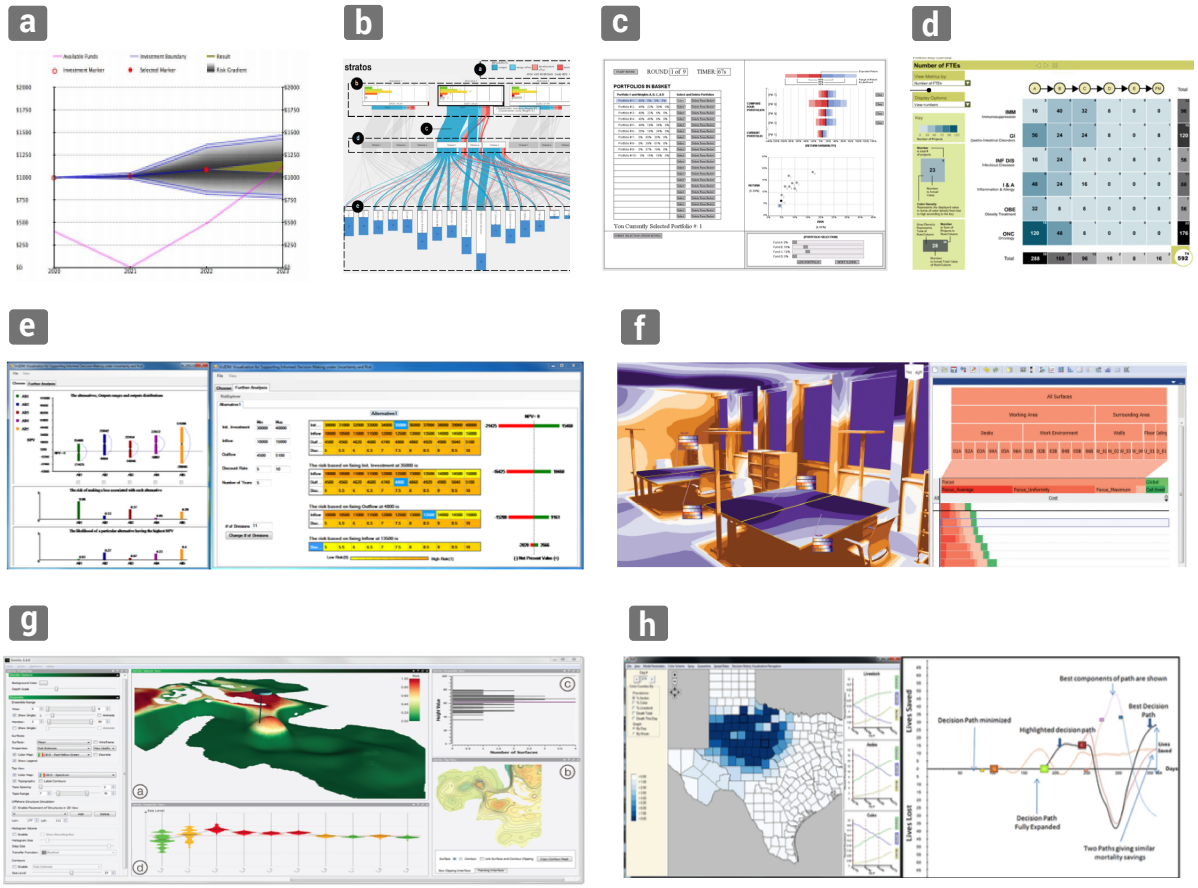


Figure 2.9: Examples of dataset-specific visualization systems that target decision-support: **a**: FinVis [405], **b**: Stratos [35], **c**: Financial Portfolio [415], **d**: Shen et al. [427], **e**: VisIDM [127], **f**: LiteVis [439], **g**: Ovis [233], and **h**: Afzal et al. [15].

and VisIDM [127] (Figure 2.9 a, c, and e). More domain-specific visualization tools for decision support exist, in areas such as lighting design [439], ocean forecast [233] and health [15] (Figure 2.9 f, g, and h).

Although most of these tools come from research, similar ones are used in industry, where visualizations can be considered critical for strategic thinking. For example, after losing millions of dollars in late drug trial failures, a large pharmaceutical company decided to use interactive visualizations to better track and facilitate decisions of "cut or go" projects in their early stages [427] (Figure 2.9 d).

2.4.1.2 Generic visualization systems for multi-attribute choice tasks

Unlike dataset-specific systems, generic decision-support visualization systems can support any multi-attribute choice task as defined in Section 2.1.2.2, by visualizing any dataset formated as a data table [361]. As generic, can be considered either systems designed to be domain independent (e.g., Value Charts) or systems that, even though they were initially designed for a particular dataset or domain (e.g., HomeFinder with houses [513] or LineUp with university rankings [201]), their design is easy to

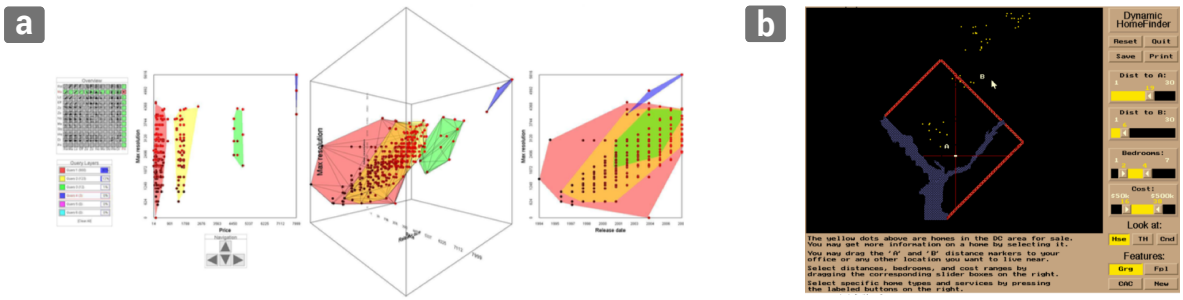


Figure 2.10: Examples of visualization systems designed for exploratory data analysis that are thought to aid multi-attribute choice due to their support for interactive querying and filtering **a**: ScatterDice [156], and **b**: HomeFinder [513].

generalize in any dataset. Generic systems are indicated with ■ in Appendix B (column "I").

Some visualization tools designed for exploratory data analysis are thought to aid multi-attribute choice due to their support for interactive querying and filtering. ScatterDice (Figure 2.10 a) is a multi-dimensional data exploration tool based on scatterplot matrices [156]. In one scenario, a user browses cameras to buy, by creating a lasso query and refining it while transitioning between scatterplots [156]. HomeFinder (Figure 2.10 b) represents data cases as dots on a map, while other attributes are represented as dynamic query widgets that can be used to progressively refine a query [513]. FilmFinder generalizes HomeFinder by changing the map into a scatterplot display [16]. Both HomeFinder and FilmFinder focus on specific domains (houses and films), but their widget-based query approach can be used with arbitrary datasets [159].

A major application area for multi-attribute choice tasks is product comparison. The vast majority of product comparison charts produced for magazines and for the Web are tables[5], with various combinations of text and visual encodings (e.g., colors, checkmarks). Similarly, a number of interactive product comparison tools, such as ManyLists [307] and FOCUS [443] (Figure 2.11 a and c), present products in a tabular visualization. Some exceptions exist. ProductExplorer (Figure 2.11 b) [396] uses parallel coordinates. SmartClient (Figure 2.11 d) [389] shows a subset of product alternatives in a scatterplot display, with a table for the remaining criteria and parallel coordinates if the users wish to apply constraints to many criteria. EZChooser (Figure 2.11 e) shows products as an image collection and encodes criteria as bargrams (i.e., histograms whose bars have been tipped over and lined up end-to-end) [514].

Some visualization systems support multi-attribute choice more explicitly, by allowing users to combine multiple attributes into a single aggregate score. Both ValueCharts [85] and LineUp [201] (Figure 2.12 a and b) initially show the choice dataset as a tabular visualization where columns can be resized to express attribute importance. The entire visualization can then be collapsed into a stacked bar chart and sorted. This approach is effectively an interactive implementation of the "weighted additive (WADD)" method described in Section 2.2.2. WeightLifter [364] (Figure 2.12 c) extends the approach

[5] As of 24 Feb 2017, the twenty top results of the search query "product comparison" on Google Images are all tables.

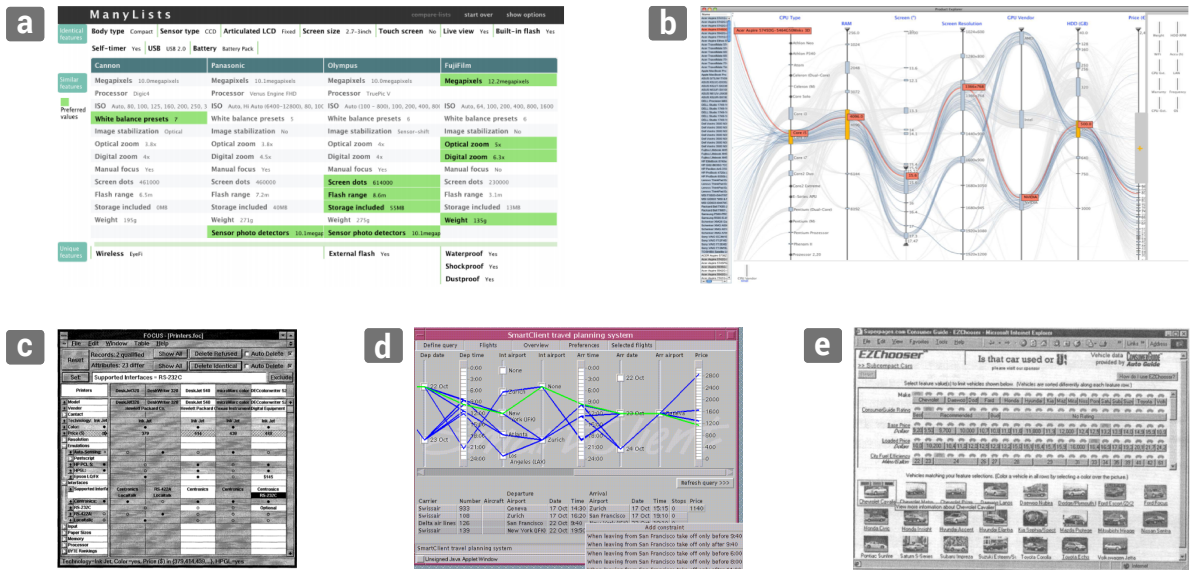


Figure 2.11: Examples of generic decision-support visualization systems for product comparison: **a**: ManyLists [307], **b**: ProductExplorer [396], **c**: FOCUS [443] **d**: SmartClient [389], and **e**: EZChooser [514].

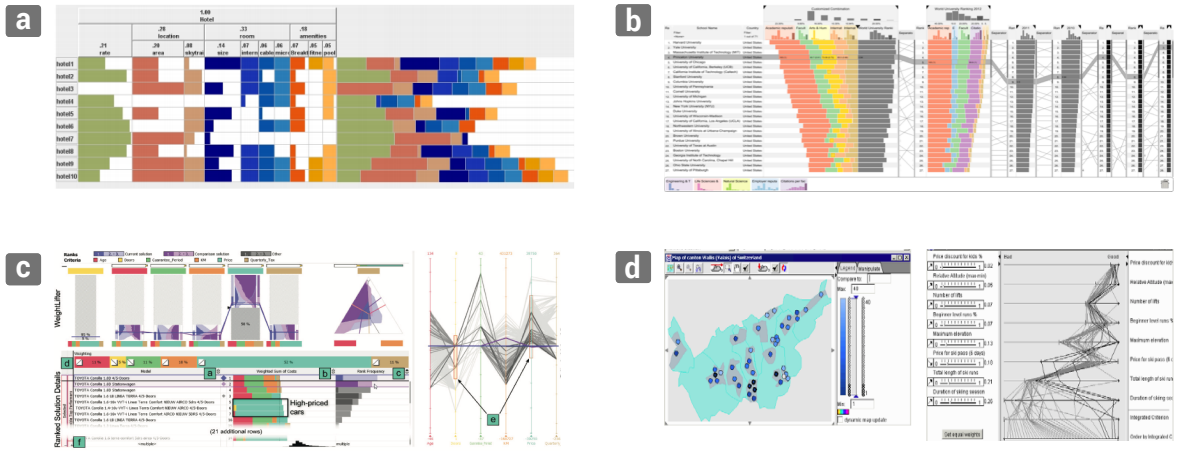


Figure 2.12: Examples of generic decision-support visualization systems that allow users to combine multiple attributes into a single aggregate score: **a**: Value Charts [51], **b**: LineUp [201], **c**: WeightLifter [364], and **d**: CommonGIS [24].

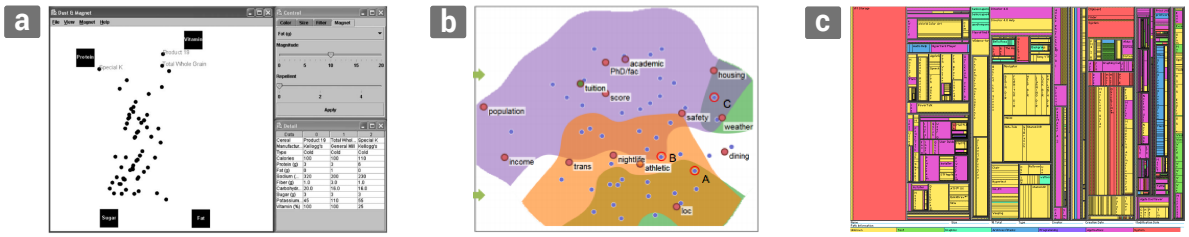


Figure 2.13: Examples of generic decision-support visualization systems with alternative designs: **a**: Dust&Magnet [525], **b**: DataContextMap [93], and **c**: AHPtreemaps [33].

by adding analytic and visualization tools such as parallel coordinates. CommonGIS also supports weighted sums, but it focuses mostly on decisions with geographical components (e.g., counties based on their need for funding, or skiing resorts) [23, 24]. It also implements a range of visualizations such as scatterplot matrices, parallel coordinates, and tabular visualizations, all linked to a map (shown in Figure 2.12 d). The explicit support for multi-attribute choice tasks using a well-structured decision strategy, such as the WADD, is likely very helpful for a decision maker. However, these systems offer limited interactions, e.g., users can not enter new data or metadata, or directly remove unwanted data cases. Moreover, no support for other decision strategies is available, e.g., "elimination by aspects (EBA)". More detailed presentation of these limitations is given in chapter 6.

As we saw, the majority of visualization tools that could support multi-attribute choice, employ traditional representations such as tabular visualizations, parallel coordinates and scatterplot matrixes. One exception is Dust & Magnet [525] (Figure 2.13 a), where queries are expressed in the form of magnets that are displayed in the same 2D space as data cases. The more a data case satisfies a query, the more it is attracted to the magnet. A scenario illustrates how a user can select cereals based on their dietary composition, by placing and moving magnets. Similarly, the Data Context Map [93] (Figure 2.13 b) , which features a scenario involving choosing a university, displays alternatives, attributes, and query results in the same unified 2D space. Another alternative representation is AHP treemaps (Figure 2.13 c) that displays attributes in a rectangular hierarchy and the choice alternatives as summary bars encoding value and importance weight [33]. The AHP treemaps system encodes each alternative with a distinct color, which, as it will be discussed in more detail in chapter 6, is likely to influence the judgment of a decision maker (e.g., a "red" encoding could be perceived as more important).

2.4.2 Evaluation of visualizations for decision support

This section reviews the methods used to evaluate the visualization systems described in the previous section. The complete list of evaluation types discussed here is also available in Appendix B in column "Evaluation" (labeled as: ■, ■, ■, ■, □, or empty white box). The color coding is explained in the legend in Appendix B and also in text in the following paragraphs. Notably, many of the visualization systems meant to support decisions do not report results from a user evaluation [15, 73, 93, 121, 156,

233, 427, 443] (indicated by the white space in the "E"column in Appendix B).

The evaluation of most visualization systems focuses on assessing whether the interaction features of the system are easy to use through qualitative studies [23, 33, 35, 127, 201, 307, 364, 389, 439, 525] (indicated by ■ and ■ in the "E"column – the difference between the two marks is whether the evaluation included decision tasks or not). Participants in these qualitative studies were only exposed to the particular visualization without having a baseline for comparison to other systems (indicated by in 'B"column). The only exception is WeightLifter [364] which briefly showed the conceptually similar LineUp system [201] to participants prior to the study (■ i.e. comparing it to another system). Moreover, to evaluate the system participants had to perform tasks that are related to understanding the data and the interactions (■), e.g., "eliminate price as a criterion"[33] or find " 3 most cited universities in the UK" [201], rather than tasks similar to the real goal of the system such as a decision task of choosing a university to apply for. Exceptions to that are Stratos system that asked participants to take the role of a project manager and choose the optimal software release plan [35]; and Dust & Magnet [525] where participants indicated which attributes of a cereal brand they consider important, and were then asked to choose a cereal brand (■). However, since the objective of these studies was to observe user interactions rather than reporting metrics of success, there is no empirical evidence that these systems indeed helped users to reach better decisions.

The few exceptions of decision support visualization systems that conducted controlled experiments (■) followed a similar methodological fashion with the qualitative studies. First, there is lack of comparative evaluations since most of them compare either variations of the same visualization [51, 110], or compare a visualization with non-visualization base cases, such as web forms [396], static numerical tables [514], or Q&A systems and textual formats [513] (■). Second, evaluations also focused on data understanding (rather than making decisions) by employing tasks such as value retrieval [51, 110, 513], range tasks [396, 513], finding extrema [51, 110, 525], finding outliers [513], and identification of patterns [513], correlations [525], and clusters [525]. Other studies involved more complex analytic tasks combining multiple low-level tasks [33, 396], but still did not ask users to make decisions. In other words, many systems suggested as being able to aid decision making are evaluated with data understanding tasks (i.e. analytic tasks).

Analytic tasks are informative when evaluating visualization tools for decision-support because good decisions require a good understanding of the relevant data. However, as we discussed in section 2.3, understanding the relevant data does not necessarily yield good decisions due to limits in human reasoning [263]. Therefore, it is important to also include actual decision tasks when evaluating visualization tools meant to support decision making.

A few controlled experiments have indeed evaluated visualization tools using multi-attribute choice tasks (■). In the evaluation of EZChooser [514], participants were asked to choose among cameras and mutual funds, and the independent measures included *decision time*, *subjective ratings* of technique preference, as well as *satisfaction* and *confidence* in one's choice. Value Charts were evaluated twice [51, 110]. In the first study, participants were asked to choose among houses, cell

phones, and tourist attractions, and the study examined the *number of insights* acquired during the decision process [51]. In the second study, participants chose among universities and restaurants, and the study examined *time*, *choice satisfaction* and *confidence* [110]. While all these studies involved actual decision making tasks, none of them used alternative visualizations as a basis of comparison (indicated with the white space in the "B"column in Appendix B) — EZChooser was compared with numerical tables, Value Charts were compared with variations of the same tool [51, 514]. Moreover, neither used objective metrics of decision quality (e.g., to validate that the chosen alternative is not dominated or is consistent with user preferences).

In the family of dataset-specific systems, there are two exceptions of controlled experiments that attempted to quantify decision quality (■): FinVis [405] and Financial Portfolio [415]. Both systems evaluated participant investments by examining the expected return of the chosen investment. However, again, none of these systems compared its design to another visualization system. Besides, as explained at the beginning of the section, the results of these dataset-specific tasks are difficult to generalize outside of the domain of financial investments. Thus, they give few insights on how and which visualizations can better support generic multi-attribute choice tasks.

2.5 Conclusion

This chapter first defined a low-level decision task named **multi-attribute choice task** and reviewed how decision theory models, normative and descriptive, explain human decisions. Later, it illustrated human decision behavior using a task-based than a model-based approach. In particular, the chapter presented a task-based taxonomy of 139 systematic errors, named cognitive biases, and identified the ones which affect choice tasks. Despite the growing interest in visualization research in cognitive biases, the main finding of this review was that there is no empirical work that either confirms or disproves the existence of cognitive biases in visual analysis.

The chapter then reviewed how multi-attribute choice tasks are supported by decision support visualizations. The review of their evaluation methodologies in Section 2.4.2 showed that the impression given from Munzner's book [347], that the effectiveness of visualizations for decision-support is not explicitly addressed, appears to be true. Although generic visualization systems for decision support are likely extremely useful, it seems that there remain important limitations in their evaluation methodologies, namely:

- very few controlled experiments;
- limited use of decision making tasks;
- lack of sensible baselines of comparison, and
- lack of metrics for decision quality.

These limitations can occur due to the fact that decisions are often subjective and have no clear ground truth, so evaluating visualizations for their ability to support decisions is difficult. Moreover, there is a lack of methodological guidance in the information visualization literature on how to conduct

such evaluations. As a result, information visualization work approach decision tasks as a high-level challenge of visual analysis [274], rather than a task that can be measured and evaluated similarly to other lower level analytic tasks (e.g., correlation, identification of trends).

The following chapter attempts to bridge this gap by considering a conceptual and methodological approach and identifies issues in evaluating visualizations for their ability to support decisions. The main challenge here is to operationalize decision making as a lower level task. Multi-attribute choice tasks will be empirically evaluated in the context of multidimensional visualizations according to various objective (e.g., accuracy, time) and subjective metrics (e.g., satisfaction, preference) that can be used to assess decision support.

Another observation from the overview of decision-support visualizations in Section 2.4.1 is that an important part of their design repeats and often combines three main types of elementary visualizations: 1) **Parallel Coordinates**, 2) **Scatterplot Matrix**, and 3) **Tabular Visualization**.

The next chapter focuses on the study of elementary multidimensional visualizations in decision support to be able to better understand more complex systems (e.g., ProductExplorer, LineUp or ScatterDice) that combine these techniques. In particular, it will investigate if the use of parallel coordinates, scatterplot matrix, and tabular visualizations can aid decision support through both a systematic analysis and an evaluation of multi-dimensional visualizations.

Chapter 3

Multidimensional Visualizations for Decision Making

The core of this dissertation is to investigate how visualizations can better support **multi-attribute choice tasks** (as defined in section 2.1.2.2). As a first step, visualization researchers need to compare the existing visualization systems in how well they can support such tasks.

The section 2.4 reviewed visualization systems specifically designed for decision support [85, 201, 364]. As previously discussed, there is very limited evidence on which system would be more effective. Most such systems have not been evaluated, and those that have been were never compared against alternative systems. The few comparative evaluations used as baselines either variants of the same system, or non-visualization formats such as web forms or numerical tables [396, 514]. A comparative evaluation of these decision support systems would be very useful; nevertheless, most of these tools are quite elaborate, often combining multiple visualizations. Before visualization researchers can study such systems, they need first to better understand the benefits and drawbacks of elementary visualization techniques.

This chapter will first articulate the link between multi-attribute choice tasks and elementary multidimensional data visualizations, by conducting a *systematic analysis* of existing **multidimensional visualizations** and the extent to which they are appropriate for multi-attribute choice tasks. Based on

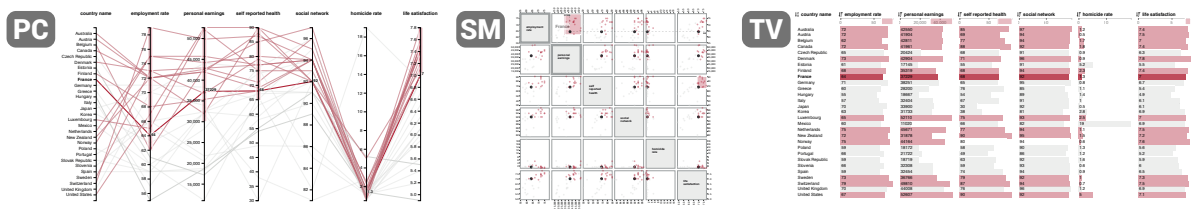


Figure 3.1: The visualizations evaluated in chapter 3: Parallel Coordinates (PC), Scatterplot Matrix (SM) and Tabular Visualization (TV).

this analysis, it will present the first comparative evaluation of three general-purpose multidimensional visualizations for their ability to support a multi-attribute choice task: parallel coordinates, scatterplot matrices, and tabular visualizations. The most important challenge of the chapter will be to outline a methodology on how to evaluate visualization techniques for decision support.

Methodological guidance on how to evaluate decisions is very critical for the information visualization field. As discussed in section 2.4.2, none of the generic visualization system designed for decision support [85, 201, 364] examined actual decision accuracy in their evaluations. Since many decision tasks are subjective and have no clear ground truth, evaluating visualizations for their ability to support decisions is difficult, and there is a lack of methodological guidance in the information visualization literature on how to do so.

In particular, the chapter will examine a methodological approach to assess decision quality by using both objective and subjective metrics. Objective decision quality metrics consist of time and accuracy. The decision accuracy is based on the consistency between the choice made and self-reported preferences for attributes. Subjective decision quality metrics consist of choice satisfaction, easiness, and confidence, as well as a novel metric of indirect confidence assessment where participants report their level of attachment in their choice over a recommendation from an expert.

Most parts of the following sections were previously published in [137]. Thus any use of “we" in this chapter refers to Evanthia Dimara, Anastasia Bezerianos and Pierre Dragicevic.

3.1 Multidimensional Visualization Approaches

Many approaches exist to visualize multidimensional datasets. Here we provide a systematic analysis of existing approaches and discuss their relation to multi-attribute choice tasks. We group them into three major families: techniques based on dimensionality reduction, non-geometric approaches, and what we call “lossless” geometric visualizations.

3.1.1 Dimensionality reduction

Some multidimensional visualizations rely on *dimensionality reduction* to collapse multiple dimensions into a smaller number of dimensions, typically two [247, 373]. Two common approaches are principal component analysis (PCA) [373] and multidimensional scaling (MDS) [247]. Although dimensionality reduction can reveal hidden structures in complex datasets and can show similar and dissimilar data cases, the resulting dimensions are often hard to interpret [420]. Furthermore, raw values are lost during the reduction process, whereas multi-attribute choice generally requires users to be able to read attribute values directly.

A related family of techniques is *dimension filtering*, which automatically removes dimensions that are either redundant or unimportant according to some criteria [522]. However, in a context of multi-attribute choice, the importance of dimensions (attributes) can rarely be deduced from the data itself as it requires personal judgment and varies across decision makers [483]. Thus, in the

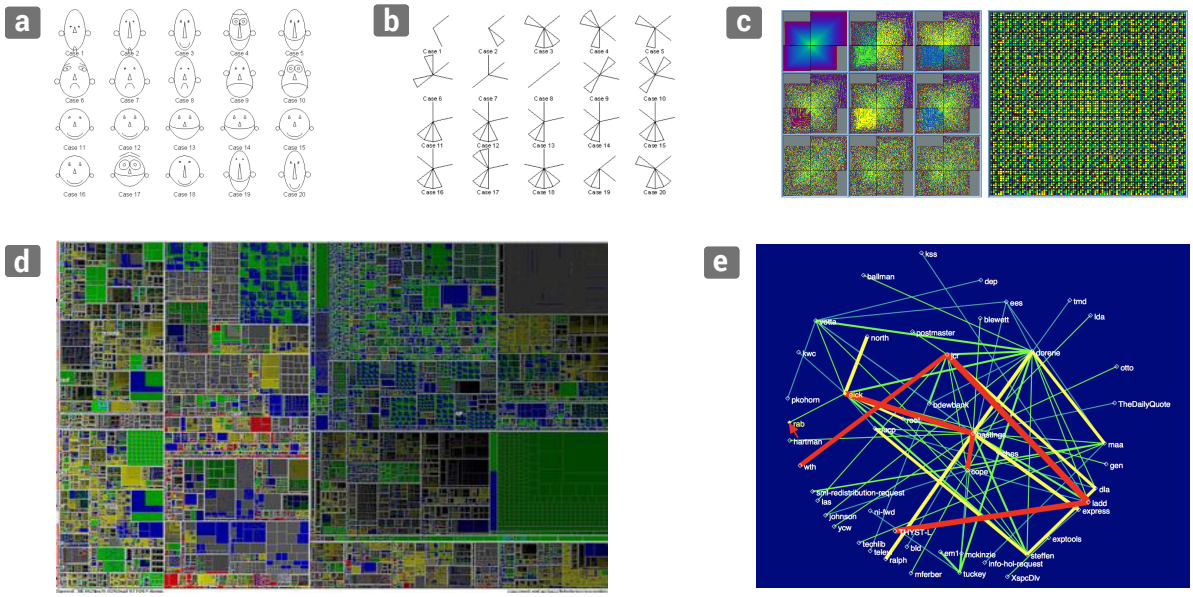


Figure 3.2: Examples of non-geometric visualization techniques: **a**: Chernoff Faces [297], **b**: StarGlyphs [297], **c**: Pixel-oriented [272], **d**: Hierarchical [428], and **e**: Graph-based [53].

absence of prior information, it seems safer to use visualizations that initially give all dimensions equal importance. Besides, the inner workings of dimensionality reduction and filtering may be hard to grasp for a general audience [420].

3.1.2 Non-Geometric visualization techniques

Keim and Kriegel [273] (also [361]) classified multidimensional visualizations into six categories, the first being *geometric projection*. Geometric projection is a broad class of techniques that encompasses both dimensionality reduction (section 3.1.1) and simpler forms of projections discussed in section 3.1.3. We discuss non-geometric approaches here.

Typical non-geometric approaches are *icon-based* techniques, where data cases are visualized side-by-side as icons or glyphs [179]. Examples include Chernoff faces [95] (Figure 3.2 a) and star glyphs [281] (Figure 3.2 b). Although icons presumably tap into our ability to visually process shapes, they can make comparisons across dimensions challenging [361], as some dimensions may be perceived as more salient than others [96, 281, 361].

In *pixel-oriented* techniques, each data case is encoded as a single colored pixel [273] (Figure 3.2 c). Examples include space filling curves [273] or spiral techniques[273].These techniques are very space-efficient and mostly useful when the number of data cases is very high. However, for common multi-attribute choice tasks, the number of data cases is rarely that high. Furthermore, color is not the most effective visual variable [105] and can impede decision making [55].

Two other categories are *hierarchical*, (Figure 3.2 d), such as Treemaps and Dimensional Stacking [273], and *graph-based* techniques [273] (Figure 3.2 e), such as Hy+[111], Margritte[482], and SeeNet

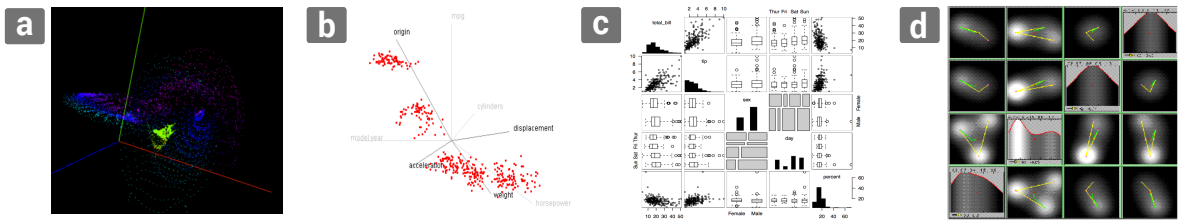


Figure 3.3: Examples of geometric visualizations based on a scatterplot: **a**: 3D scatterplots [286], **b**: Star Coordinates [267], **c**: Generalized SM[157], and **d**: HyperSlice SM [231].

[53]. These techniques assume the existence of structural relationships between attributes that may not be available in multi-attribute choice situations. Finally, *hybrid* techniques combine multiple visualizations either in-place or side-by-side [361]. Although combining different approaches can be powerful, the strengths and weaknesses of elementary visualization techniques need to be better understood before we know how to combine them effectively.

3.1.3 Lossless geometric projection

Keim and Kriegel's taxonomy [273] can be refined by splitting *geometric projection* techniques into *lossy* and *lossless*. As we discussed, visualizations employing dimensionality reduction are lossy because raw values are lost and cannot be retrieved by looking at the visualization. For example, an MDS projection can lay out cameras on a 2D space so that similar cameras are close to each other [247], but users cannot read the price or resolution of cameras unless separate detail-on-demand techniques are provided. In contrast, in a lossless projection, *any attribute value from any data case can be visually retrieved without interactions beyond basic scrolling and panning operations*. Thus, although in practice lossless projections may require interaction if the dataset is too large to fit the screen, in principle no interaction is required if the display is sufficiently large.

Lossless geometric projection approaches employ simple visual encodings and encompass some of the most commonly used multidimensional visualization techniques [347, 515].

A table dataset with two dimensions can be visualized losslessly with a *2D scatterplot*. 2D scatterplots can be extended to more dimensions by employing either higher-dimensional scatterplots (e.g., 3D scatterplots [286] in Figure 3.3 a) or star coordinates [267] (Figure 3.3 b). However, since the location of each data point on the display encodes a vector sum, both techniques are lossy. A lossless alternative involves creating 2D scatterplots for every pair of dimensions and arranging them in a *scatterplot matrix* [157]. Many variations of scatterplot matrices have been proposed, including versions that use color encodings [180, 479], or extensions that support categorical data [157, 246] (Figure 3.3 c) or continuous multidimensional functions [509] (Figure 3.3 d).

Another classic lossless geometric projection technique is the *parallel coordinates plot*, where dimensions are parallel axes, and data cases are polylines that intersect the axes at their corresponding values [248]. Variations of parallel coordinates exist that are circular [231] (Figure 3.4 a), hierarchical

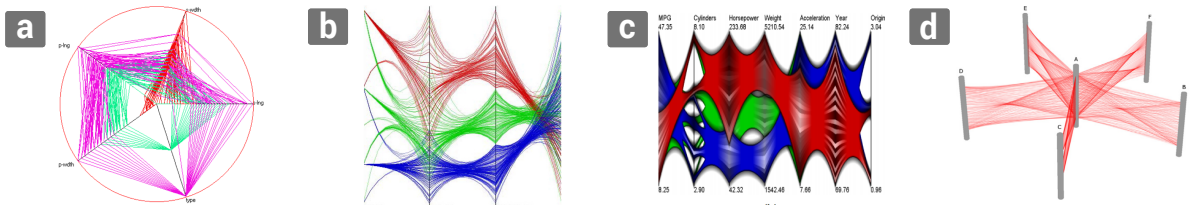


Figure 3.4: Examples of geometric visualizations based on parallel coordinates: **a**: PC Circular [231], **b**: PC Bundled[257], **c**: PC curved [329], and **d**: PC 3D [257].



Figure 3.5: Examples of geometric visualizations based on a tabular visulization: **a**: TableLens [312] and **b**: Bertifier [374].

[178], bundled [533] (Figure 3.4 b), curved [22] (Figure 3.4 c) or use 2D-3D layouts [258, 502] (Figure 3.4 d). However, according to a recent survey [257], there is not enough empirical support to suggest that the alternative configurations outperform the original representation. A hybrid technique has also been proposed that combines parallel coordinates with scatterplot matrices [485].

A third lossless technique is the *tabular visualization*, i.e., a numerical table whose cell values are encoded visually [374, 512]. Common encodings include length (bars) [58, 374, 391] (Figure 3.5 a), luminosity or hue (i.e., shaded cells) [374, 512], and area (e.g., circles) [58, 374, 429] (Figure 3.5 b).

Although tabular visualizations are not as popular as scatterplot matrices and parallel coordinates in information visualization [347, 515], they have long been used in some scientific circles [58, 374, 512] and have been occasionally promoted within infovis because of their flexibility and their efficient use of space [130, 226, 374, 429]. Tabular visualizations are supported to some extent by most modern spreadsheet software through a "conditional formatting" feature, where numerical values are generally displayed on top of the encodings [374].

Stacked bar charts and *grouped (or clustered) bar charts* are analogous to tabular visualizations that use bars to encode values, except bars are stacked or displayed next to each other instead of being aligned. Although stacked and grouped layouts are commonly used in statistical charts, studies have suggested that the aligned layout of tabular visualizations has perceptual benefits [105, 221, 456, 521]. Furthermore, stacked and grouped bar charts need to encode bars of the same category with color, which limits their scalability as multidimensional visualizations due to human limitations in color discrimination [495].

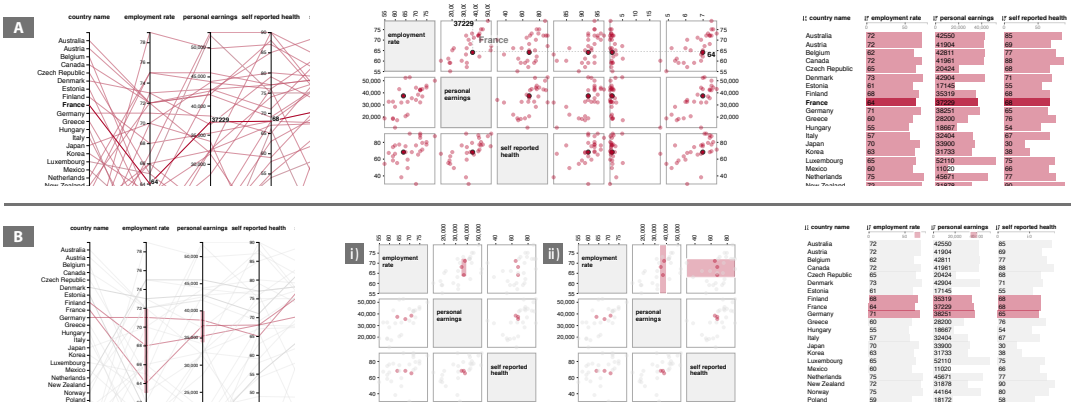


Figure 3.6: A) Single data case “France” highlighted on hovering. B) Range selection on two dimensions. The three data cases at the intersection of the selected ranges (63-72 for “Employment rate", and 340000-40000 for “Personal earnings") are “France”, “Germany” and “Finland”.

3.1.4 Evaluations

Previous studies suggest that 2D scatterplots outperform bivariate parallel coordinates for correlation tasks [303], and that scatterplots embedded within parallel coordinates outperform parallel coordinates alone for cluster detection [234]. A more recent study [291] compared parallel coordinates with three simplified forms of scatterplot matrices (where only a subset of the plots is shown) for basic value retrieval tasks, and found that one of the simplified forms outperformed parallel coordinates. However, it remains unclear whether complete scatterplot matrices (i.e., that include all $n(n-1)$ pairs of dimensions) would also outperform parallel coordinates in value retrieval if screen real-estate is controlled for. At the same time, simplified scatterplot matrices hide most bivariate relationships, and thus, may not be as suitable for overview tasks such as identifying highly correlated dimensions.

Evaluation of multidimensional visualization techniques is still in its infancy. We know little about how elementary multidimensional visualizations compare in terms of elementary analytic tasks, and even less so in terms of how they support decision tasks. In particular, we do not know of any study that examines decision tasks, and little or no comparison based on basic analytic ones.

3.2 Technique Design

As we discussed in section 3.1, we focus our evaluation on lossless geometric projection techniques, as they are widely used, they support attribute value retrieval, and they can accommodate a range of multi-attribute choice datasets (e.g., they do not require attributes to be hierarchically organized). In particular, we examine three, commonly used, elementary lossless geometric projection visualizations: the parallel coordinates (PC), the scatterplot matrix (SM), and the tabular visualization (TV).

Our evaluation methodology relies on two major principles: *(i)* include all features that are considered standard for each visualization, *(ii)* keep the visualizations as comparable as possible through a

consistent visual design, a consistent interaction design, and by having all interactions present the same amount of information across visualizations.

3.2.1 Visual encodings

Our implementation employs the most commonly used visual marks to represent data cases: polylines for PC, dots for SM, and bars for TV. We keep the visual design as consistent as possible across the techniques, to facilitate comparison. For example, visual marks share the same color across all techniques (translucent red by default, or translucent gray when outside a range selection), while decorations (e.g. axes, fonts) are consistently displayed in gray or black. The three techniques occupy similar vertical screen space, although the total area of SM is smaller due to its square aspect ratio that is not adapted to the typical landscape orientation of computer displays. More details are given next.

Parallel Coordinates (PC). We use the original representation introduced by Inselberg in 1960 [248]: a polylines diagram where the dimensions are represented as parallel axes and the data cases as polylines that intersect the axes at their corresponding values [257] (see PC in Figure 3.1). This representation is considered standard in several infovis textbooks and surveys [347, 515]. As we saw in the Background section, many variations exist, but there is not enough empirical evidence that they outperform the original layout [257].

Scatterplot Matrix (SM). We use the full matrix, defined by Emerson et al. [157] as *"a grid of scatterplots showing the bivariate relationships between all pairs of variables in a multivariate data set"* (see SM in Figure 3.1). As we have seen, simplified forms of scatterplot matrices exist that only show a subset of plots[291], but the complete scatterplot matrix (either square or triangular) has the advantage of showing all attribute pairs and is by far the most widely used [46, 87, 104, 156, 157, 299, 303, 347, 515].

Tabular Visulization (TV). We encode cell values by length (bars) [58, 374] (see TV in Figure 3.1). Although other encodings exist (see section 3.1.3), we followed the Table Lens example [391] of choosing length, because it is more accurately perceived than other visual encodings [104], and because it is commonly used in tabular visualizations for decision support [23, 85, 201, 364]. We also display the numerical values on top of the bars, as is usually done in current spreadsheet software through the "conditional formatting" feature [374].

3.2.2 Interaction techniques

Interaction is essential for analytic and visual exploration tasks, and likely also for decision making tasks. We chose to support three types of interactions which are either considered standard for at least two of the techniques, or have proven useful in decision making tools:

Highlighting of individual data cases with *linking* and *details-on-demand* to support value retrieval across all criteria [85, 201, 364, 396] (see Figure 3.6 A). Single data cases can be highlighted by hovering over a data case, which changes the opacity of the entire data case (line, dot or bar depending on the visualization) from the default 40% to 100%. Hovering over a data case or a dimension axis displays the precise values of the data case with tool-tips. Brushing and linking is commonly used in all techniques

to highlight one or several data cases so as to help users relate their values across dimensions [430] or of the same row. In SM the data case highlighted in all plots (linking) assists users to relate the different views [52, 156, 347, 515].

Range selection on one or more dimensions to support dynamic filtering and queries [85, 201, 307, 389, 396] (see Figure 3.6 B). This results in graying out all data cases outside the selection. If range selection is performed across multiple dimensions their intersection is shown, i.e., data cases that simultaneously fulfill the selected ranges for each dimension. Range selection in PC and TV is performed by brushing an axis, which in TV is located below the column titles [198]. Range selection is slightly different in SM given the bivariate nature of scatterplots (Figure 3.6 i and ii). Instead of brushing individual axes, users draw selection rectangles inside the scatterplots. This effectively selects two ranges at the same time (one for each dimension of the scatterplot). All range selections are re-sizable and drag-able through handles that appear on hover. While drawing or adjusting a range, the value of the range limits is displayed on the corresponding axes (not visible in the figures).

Dimension reordering to allow users to sort attributes by preference [156, 201, 396]. Rearrangement of dimensions brings together the ones relevant to the task. In our implementation, it is performed by dragging axis titles for all techniques. Reordering is fairly common in PC [430], and SM occasionally includes methods for manually or automatically reordering dimensions [156]. Unlike PC, though, in SM all possible pairs of dimensions are shown and thus reordering is not essential to perform side-by-side comparisons of dimensions. Reordering is also considered essential in TV, and research prototypes typically support manual reordering not only of columns, but also of rows [130, 226, 374, 429]. Thus we also allow manual reordering of rows (data cases) in TV. Reordering data cases is impossible in PC and SM since the position of visual marks is determined by the data. Most research prototypes of tabular visualizations also support automatic reordering of rows or columns based on similarity [130, 226, 374, 429], but we considered these features as too advanced for a comparison of elementary visualization techniques. Nevertheless, column sorting (a simple form of reordering) is a central feature of all commercial spreadsheet software tools, so we decided to include it in TV as well (both ascending and descending).

3.3 Experiment

Our goal is to explore how to evaluate elementary multidimensional visualizations for their ability to support decisions. To this end, we compare PC, SM and TV (see Figure 3.1) according to how well they can support *i) basic data exploration*, by giving participants analytic tasks; and *ii) decision making*, by giving participants multi-attribute choice tasks. The reason behind this dual evaluation is that elementary analytic tasks can be thought as necessary, as we will explain in section 3.3.2, but not necessarily sufficient components of multi-attribute choice tasks. By starting with basic tasks, we can train participants in reading and interacting with the visualizations before they proceed with the decision task. Doing so also allows us to ensure that they properly understood the techniques when

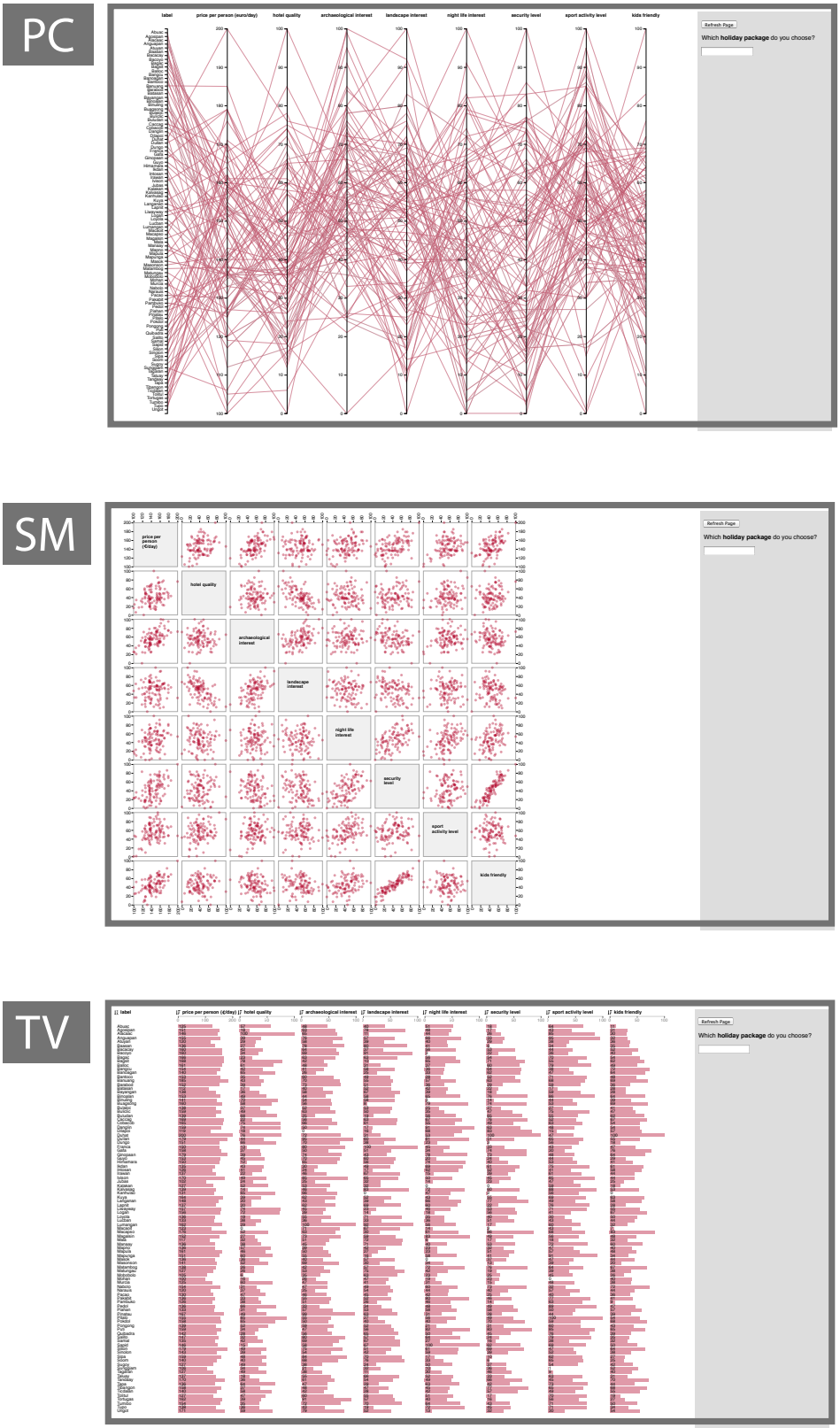


Figure 3.7: Experiment Stimuli for the decision task (“Which holiday package do you choose?”). Dataset of 100 holiday packages.

they performed the choice task, thus eliminating potential confounds (e.g., a technique yielding poor decisions because participants did not know how to use it). Finally, a dual evaluation may uncover potentially interesting interactions between a technique's ability to support analytic tasks and its ability to support decision tasks. Again, there is currently little empirical data we can draw from on how the three techniques compare even for basic analytic tasks.

3.3.1 Research questions

Prior to data collection we framed the following research questions:

Q1 [ACCURACY] *How do the three techniques compare in terms of accuracy in **a**) analytic tasks and in **b**) decision tasks?*

Q2 [TIME-ON-TASK] *How do the three techniques compare in terms of speed in analytic tasks?*

Q3 [SUBJECTIVE PREFERENCE] *Which technique people prefer overall for **a**) analytic tasks **b**) decision tasks?*

Q4 [SUBJECTIVE CHOICE ASSESSMENT] *How do the three techniques compare in decision tasks in terms of choice **a**) satisfaction, **b**) confidence, **c**) easiness, and **d**) attachment?*

We did not initially consider time for decision tasks as part of our initial research questions, as we wanted to focus on accuracy and subjective satisfaction. All metrics are described in sections 3.3.9 and 3.3.10.

3.3.2 Tasks

We used three *analytic* tasks inspired from standard visualization taxonomies [20, 402] and one *decision* task:

Value Retrieval. The task consisted of identifying the alternative having a certain attribute value and finding the value of another of its attributes [291]. Reading individual attribute values is likely very common in multi-attribute choice tasks. Value retrieval is also often considered as a building block of tasks like "find extrema" or "sorting" [19, 291], that are both common in decision making [201, 364].

Range. The task consisted of finding how many alternatives have their attribute X within a given range, and their attribute Y within another given range. This task is analogous to the "compute derived value" task [19]. It is likely involved in multi-attribute choice tasks when filtering alternatives, including when discarding unattractive options that do not match the decision makers' preferences and constraints.

Correlation. The task consisted of finding the pair of attributes that have the strongest correlation. This is an overview task, in contrast to correlation tasks that require to estimate the correlation of a single pair of dimensions [19, 402] or to compare the correlation between two pairs of dimensions [394]. Identifying strong correlations can be involved in decision tasks where relations and trade-off comparisons between pairs of attributes are important [364]. For example, detecting a high correlation

between two attributes such as quality and price may lead to a search for outliers which are particularly "good deals".

Decision. The task was a multi-attribute choice task as defined in section 2.1.2.2. It consisted of finding the best alternative (in terms of subjective desirability) among a fixed set of alternatives (see Figure 3.7).

3.3.3 Datasets and task generation

We used three different datasets:

Training. For the training, we used a dataset of country indicators from www.oecdbetterlifeindex.org, from which we selected 29 countries and 6 dimensions (e.g., life satisfaction, homicide rate, etc.).

Analytic tasks. For the analytic tasks, we used 18 synthetically generated datasets of student grades, containing 100 data cases each (students) and 8 dimensions (grades for different subjects such as English, math, biology, etc.). Grades were between 0 and 100.

Decision task. For the decision task, we used 3 synthetically generated datasets of holiday packages, containing 100 data cases each (holiday packages) and 8 dimensions: price per person (euro/day), hotel quality, archaeological interest, landscape interest, night life interest, security level, sport activity level, and kids friendly. Prices were between 100€ and 200€. All other dimensions were ratings from 0 to 100. Package names were generated using the City & Town Name Generator (www.mithrilandmages.com/utilities/CityNames.php).

For both the analytic and the decision datasets, correlated data was generated by sampling from random positive definite covariance matrices using the R packages clusterGeneration and MASS. Datasets were regenerated until the difference between the highest and the second highest absolute correlation was at least 0.3. For the analytic dataset, the highest correlation additionally had to be positive, and its two attributes had to be separated by at least a column. For the holiday dataset, price had to be positively correlated with all other dimensions.

Each analytic dataset yielded a *correlation task*. In addition, we generated a *value retrieval* task by randomly choosing a data case and two attributes (one to locate the data case, one to read the value), such that *i)* the attributes are separated by at least a column, and *ii)* the value of the attribute used to locate the data case is separated from the closest value by at least 0.02 (axes normalized between 0 and 1). We also generated one *range task* per dataset by choosing two random attributes and value ranges such that *i)* the two attributes are separated by at least a column, *ii)* each endpoint of each range is separated from the closest value by at least 0.02, *iii)* range widths are between 0.1 and 0.8, *iv)* each range contains 1 to 5 data cases, and *v)* the intersection between the two ranges contains fewer data cases than either range alone.

3.3.4 Apparatus

We used a 1920x1080 resolution screen, with mouse and keyboard as input. The visualization software was implemented in D3, and questionnaires (see section 3.3.8) were shown on Google web forms.

3.3.5 Techniques

The three techniques (PC, SM and TV) are illustrated in Figure 3.7 and explained in detail in section 3.2. Each visualization entirely filled the vertical display space, and for PC and TV, the horizontal display space. Each visualization could accommodate the seven attributes without the need for scrolling, and with legible fonts.

3.3.6 Participants

We recruited 21 participants (6 female) among students, engineers and researchers working in computer science, with a mean self-reported experience in data visualization of 6.0/10 (range 2–9, σ : 1.66).

3.3.7 Experiment design

We used a within-subjects design with independent factor the *visualization technique* (PC, SM and TV). The experiment was divided into two sessions. In the *analytic* session, participants performed the three analytic tasks in a fixed order: *four* trials of the value retrieval task, then *four* trials of the range task, then *two* trials of the correlation task, using the "student grades" dataset described in subsection 3.3.3. During pilot testing the correlation task took much longer, so we decided to only include two trials to keep the experiment time manageable. Two training trials were performed before each new task. The presentation order for visualization technique was counterbalanced using a latin square.

In the *decision* session participants performed one decision task per technique, using the "holiday packages" dataset described in subsection 3.3.3. This dataset was generated in a similar manner as the analytic dataset, but used different random values as well as different names for attributes and data cases in order to prevent the analytic session from influencing decisions and strategies used in the decision session. The order of the decision tasks was fixed while the technique presentation order followed that of the analytic session, effectively counterbalancing the dataset/technique pairing.

3.3.8 Procedure

We conducted a pilot study to ensure the clarity of the instructions and estimate task time. Our final experiment lasted on average 1.4 hours (ranging from 1.1 to 1.7 hours) and consisted of the following steps.

Technique Training: At the beginning of the experiment and before each change of technique, participants were shown, in a paper, a table representation of a minimalistic dataset (5 data cases) next to the introduced technique. The experimenter then explained how to read the visualization, by marking in different color the corresponding encodings (lines, dots or rows) for a single data point, and indicating with dashed lines how to project them to the axis to read values. For the SM the experimenter further explained its arrangement, noting that they can see two attributes per scatterplot and symmetrical plots along the diagonal present the same information. Participants were next shown the interactive version with the training dataset described in subsection 3.3.3 (see Figs 3.6). For each

interaction (highlighting, range selection and reordering), the experimenter explained the interaction and invited the participants to try on their own. A summary of all instructions was provided on a cheat-sheet paper that was visible by participants during the experiment.

Task Training: After technique training, participants moved into performing the analytic tasks as described previously. Each type of task was preceded by two training trials, one performed by the experimenter to illustrate the task, and one by the participant. When participants indicated they had understood the task, they moved on to performing the experimental trials without assistance. Participants typed their answer (value in the retrieval task, number of items in the range task, and pair of dimensions in the correlation) in a text field provided to the right of the screen (see Figure 3.7). At the end, participants filled in a technique preference questionnaire described in 3.3.10.1.

Decision Task: After performing all analytic tasks with all techniques, participants were told they would now use the techniques to make a personal choice. They were asked to imagine planning their vacations and looking for the ideal holiday package. The meaning of each of the attributes of the holiday dataset was explained, and they were informed that they would see a different set of holiday packages each time. Participants conveyed their choice by copying the package's name in a text field provided to the right side of the screen.

As we will explain in subsubsection 3.3.9.1, before the first, and after each decision task (4× total) participants filled-in a questionnaire to indicate which attributes they consider important. After each task, they also filled in a questionnaire to assess their satisfaction with their choice. At the very end of all decision tasks, participants filled in a questionnaire on their overall technique preference for decision tasks.

3.3.9 Objective performance metrics

We collected accuracy and time-on-task measures for both tasks. Accuracy in particular is a challenging measure to define in decision making, an inherently subjective task. Details are provided next.

3.3.9.1 Accuracy

For all tasks, we used a normalized measure of *accuracy* ranging from 0 to 1, with 1 being a fully correct answer. We used continuous measures whenever possible to maximize statistical power.

Analytic tasks: In the *value retrieval* task, where participants needed to find the value of an attribute, we gave a binary score (1 = correct, 0 = incorrect). A partially correct answer was difficult to define as values close to the correct value were often shared with other items. Thus there was no way to determine if an incorrect response was due to an incorrectly identified data case or due to a misread value. In the *range* task, where participants needed to count data cases, accuracy was defined $a\ 1{-}¼|correct-response|$. All range tasks involved from 1 to 5 items, thus the normalizing term $5-1\ 4$. In the *correlation* task, accuracy was defined as $a\ 1-|correct-response|$, where *correct* was the highest correlation in the dataset, and *response* was the correlation between the two attributes given as a response.

Decision tasks: There is no simple way to define the accuracy of a multi-attribute choice task, given its subjective nature. Although dominance is one such measure as discussed in chapter 2.2.1, the selection of a dominated alternative is unlikely in our dataset given the number of alternatives and attributes. We thus decided to use as an indicative measure of accuracy the *consistency between the choice made by a participant and her self-reported preferences.* As mentioned before, participants rated the *importance* of each of the 8 holiday package attributes between 0 and 10. For example, some may consider as of priority the destinations of high archaeological interest whereas others may search for destinations with lively nightlife or exceptional nature. They also indicated the *direction* of their preference, i.e., whether they prefer the attribute to be high or low. For example, a holiday package with lots of physical activity can be perceived as desirable by an athletic person but undesirable by someone with reduced mobility. As preferences may evolve during the session, the questionnaire was administered before and after each decision task (4× total).

Based on this data, we can roughly estimate how desirable each alternative should be using a weighted sum approach [466]. For each participant and decision dataset, we compute a *desirability score* per alternative as follows: for each attribute, *i)* divide its value by the maximum allowed value, *ii)* if the user's preference is toward small values, replace the value with $1 - value$, *iii)* multiply the value by the attribute's importance obtained from the questionnaire (0–10). Once done, sum up all attribute values to obtain a desirability score d for that particular alternative. We repeat the process for all alternatives, then normalize all d scores between 0 and 1. Thus, the "optimal" alternative in the dataset has a d of 1 while the worst one has a d of 0.

Desirability scores can be computed using the preferences elicited either before the decision task (d_{pre}), or after the task (d_{post}). Since preferences can evolve while exploring options, d_{post} may seem more indicative of the "true" desirability. However, a participant may also update their preferences *after* the choice was made, e.g., as a way of rationalizing their choice. Thus, we consider both d_{pre} and d_{post} and define the accuracy of a decision task as $a\ \max(d_{pre}, d_{post})$, with d_{pre} and d_{post} being the desirability scores of the chosen alternative. This score is an approximation and is not meant to capture decision quality perfectly. The elicited preferences may not be completely reliable, and cannot fully capture the complexity of choice criteria (i.e., someone may want an attribute to be neither too high nor too low). However, if a visualization happens to be misleading or particularly hard to use, we can expect participants to make choices that are clearly inconsistent with their preferences, thus yielding an abnormally low precision score.

3.3.9.2 Time-on-task

Analytic tasks: We consider the time participants took to complete each analytic task, from the moment the task page is displayed to pressing the ENTER key after giving the answer.

Decision tasks: We did not consider completion times for decision tasks in our planned analysis, but considered including them in posthoc analyses. Time was measured from the moment the decision dataset was shown, to when participants pressed ENTER to confirm their choice.

3.3.10 Subjective metrics

We considered two types of subjective metrics: technique preference (for analytic and decision tasks) and choice assessment (for decision tasks). All responses were reported on an 11-point scale, from 0 to 10.

3.3.10.1 Technique Preference

We asked participants to rate the techniques based on overall preference.

Analytic tasks: After completing all analytic tasks (value, range, and correlation) with all techniques, participants were asked to rate how easy and helpful they found each technique. They were orally instructed not to focus on a specific analytic task but on their overall experience. They were also given the option to justify their ratings.

Decision tasks: Similarly, after completing all decision tasks with all techniques, participants were asked to rate how easy and helpful they found each technique for choosing a holiday package.

3.3.10.2 Choice assessment

After completing each decision task (one per technique) and before the next preference elicitation questionnaire, participants evaluated the choice they just made according to the following criteria:

- *satisfaction*: Participants were asked to what extent they are satisfied with their choice ranging from "not satisfied at all" to "very satisfied";
- *confidence*: They were asked to what extent they are confident about their choice ranging from "not confident at all" to "very confident";
- *easiness*: They were asked to what extent they consider this choice as easy to make ranging from "very difficult" to "very easy";
- *attachment*: Participants were asked to imagine that an automatic recommender system could suggest another choice from the dataset taking into account their preferences, and were asked whether they would switch to this choice ranging from "I would definitely stick to my initial choice" to "I would definitely switch".

The first three subjective metrics are often used in decision support tool evaluations [51, 201, 525]. They are meant to complement the objective accuracy metrics described previously, by explicitly asking the participants to evaluate their choice. The fourth metric (attachment) is based on Chernev [94], and provides a more indirect way of asking participants to evaluate their choice. Chernev used this metric as the primary dependent variable in a decision making study involving the evaluation of consumer choices, assuming that participants who are confident in their choice will have less propensity to switch [94].

3.4 Results

We analyze, report and interpret all our inferential statistics using interval estimation [140]. Experimental stimuli, data and analyses are available at `http://www.aviz.fr/dm`.

3.4.1 Planned analyses

In this section, we focus on the analyses planned before data was collected. Each subsection corresponds to one of our research questions stated in subsection 3.3.1, with the same notation **Qx**. All differences between techniques are computed within-subjects (paired samples).

3.4.1.1 Q1 – Accuracy

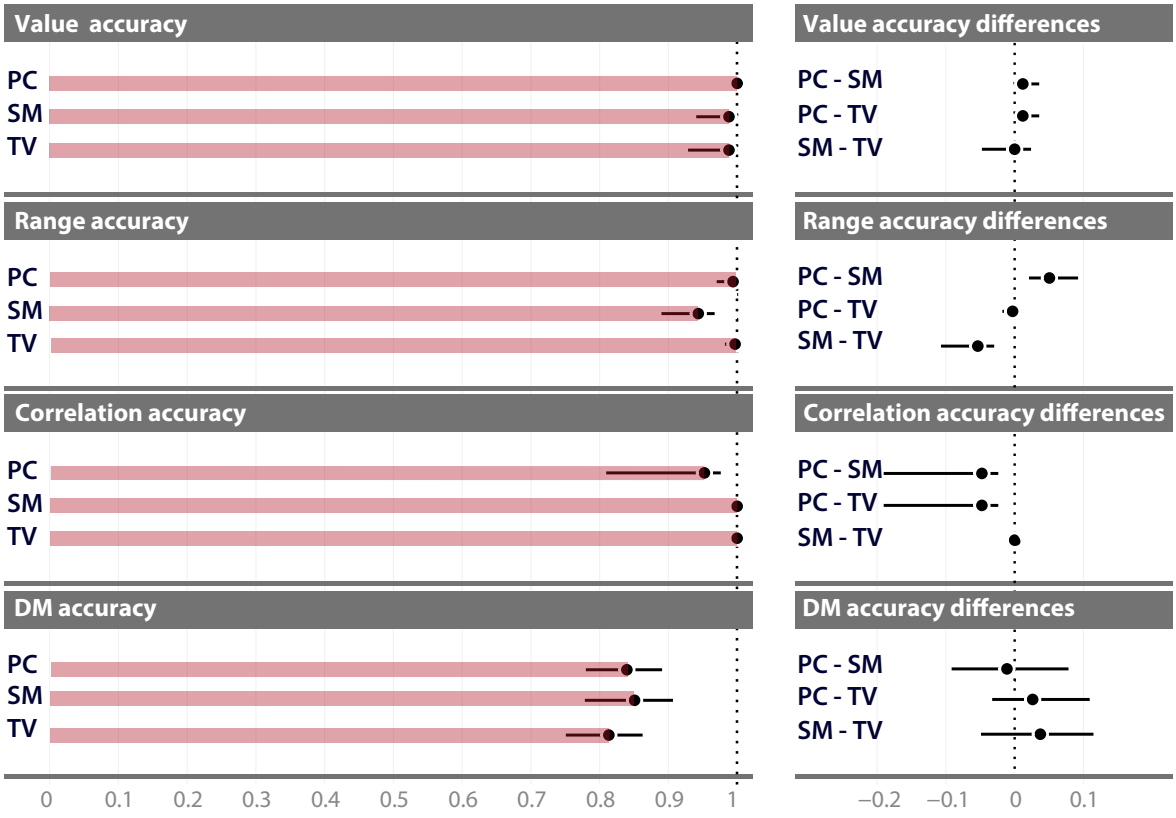


Figure 3.8: *Left:* Mean accuracy scores achieved for the three analytic tasks and the decision making (DM) task, using the parallel coordinates plot (PC), the scatterplot matrix (SM), and the tabular visualization (TV). *Right:* Mean differences in accuracy scores between each pair of techniques — a positive value indicates that the left technique outperforms the technique on the right. All error bars are 95% CIs (n=21).

Results for accuracy are reported in Figure 4.6. Each of the four horizontal panels shows the results for one type of task. The top three panels report accuracies for the analytic tasks (value retrieval,

range and correlation), while the bottom panel reports accuracies for the decision task. The bar charts on the left show the *mean accuracy* score for each technique, while the dot plots on the right show the *mean differences in accuracy* between techniques. A positive value (to the right of the zero axis) indicates that the left technique outperforms the right one. For each mean, a point estimate is reported together with a 95% confidence interval (CI) indicating the range of plausible values for the population mean [140]. All confidence intervals are 95% BCa bootstrap CIs [277].

Q1a. We can see that participants achieved a perfect or close-to-perfect accuracy score in almost all analytic tasks (mean scores: Value PC 100%, SM+TV 99%; Range PC 99%, SM 94%, TV 100%; Correlation PC 0.95%, SM + TV 100%). The two exceptions are the range task carried out with SM, and the correlation task performed with PC. In both cases, participants were reliably less accurate than with the other techniques, but the differences are relatively small. This means that participants followed the instruction to be as accurate as possible, and completion times (analyzed thereafter) should give a good indication of overall performance with analytic tasks.

Q1b. For the multi-attribute choice tasks (DM), participants were on average fairly accurate with all techniques in terms of how consistent their choices were with their self-reported preferences (mean score: 81-84% for all techniques). That said, no technique yielded a perfect or close-to-perfect average accuracy score, meaning that participants rarely made an "optimal" choice regardless of which technique they were using. Interestingly, there is no sign of a clear difference in accuracy between the three techniques.

3.4.1.2 Q2 – Time-on-Task

Figures 3.9 and 3.10 present the average amount of time spent by participants in each analytic task. As before, mean observations (in seconds) are reported to the left. This time, raw measurements were log-transformed to correct for positive skewness and reduce the influence of extreme observations, and were then antilogged at the end of the analysis [140]. As a consequence, all reported means are geometric means, and differences between techniques are expressed as ratios of mean completion times (reported to the right). A value to the left of the x=1 axis means that the numerator technique is faster than the denominator technique. All confidence intervals are exact confidence intervals for the normal distribution, computed on the logged observations.

Q2. The top horizontal panel in figures 3.9 and 3.10 report completion times and differences respectively for the *value retrieval* task. The figure provides strong evidence that participants were much slower on average (almost twice as slow) with SM than the other two techniques, which are comparable in speed (mean times in sec: PC 18, SM 34, TV 19).

The results are similar for the *range* task (second panel), with SM being again almost twice as slow as the other two techniques (mean times in sec: PC 32, SM 57, TV 30). TV is possibly slightly faster than PC (ratio PC/TV of 1.1, 95% CI [0.99,1.2]).

The results are very different for the *correlation* task. SM is remarkably fast: about 9 times faster than PC and 4 times faster than TV (mean times in sec: PC 100, SM 12, TV 50). Here PC is clearly the

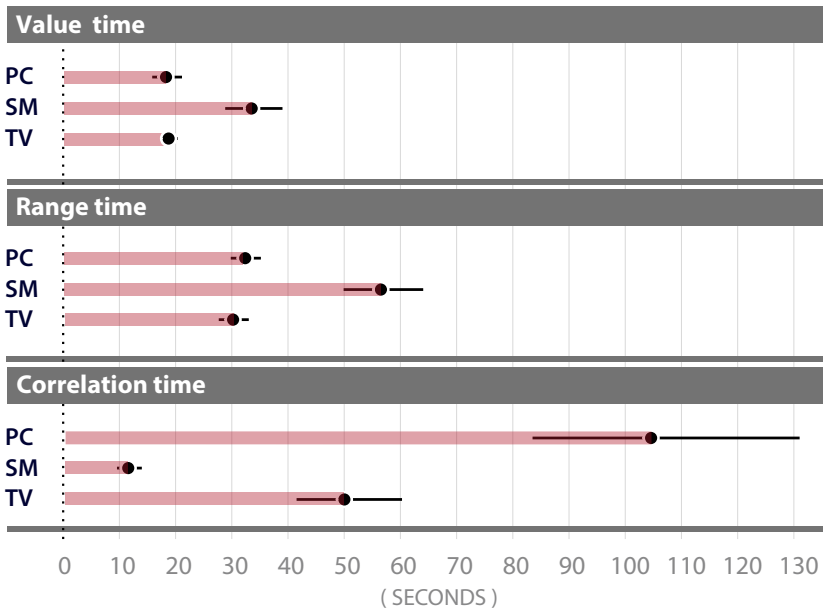


Figure 3.9: Average time (in seconds) spent on each analytic task for techniques PC, SM and TV. All error bars are 95% CIs (n=21).

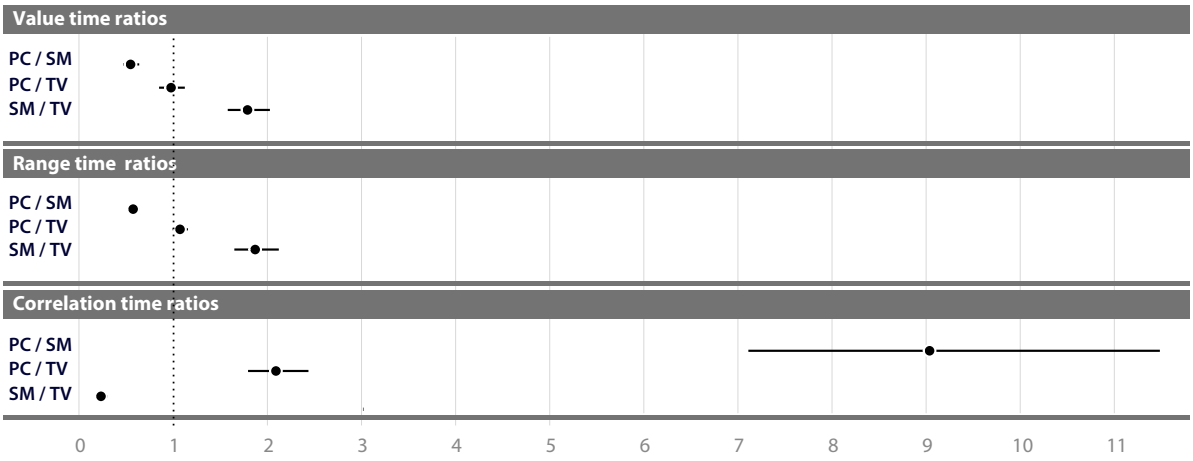


Figure 3.10: Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21).

slowest, with TV being about twice as fast as PC.

3.4.1.3 Q3 – Subjective Preference

Figure 3.11 presents mean participant ratings, in terms of how easy and helpful they felt the techniques was when carrying out analytic tasks (top panel) and decision making tasks (bottom panel). On the difference plots, a positive value indicates that the left technique is on average preferred to the

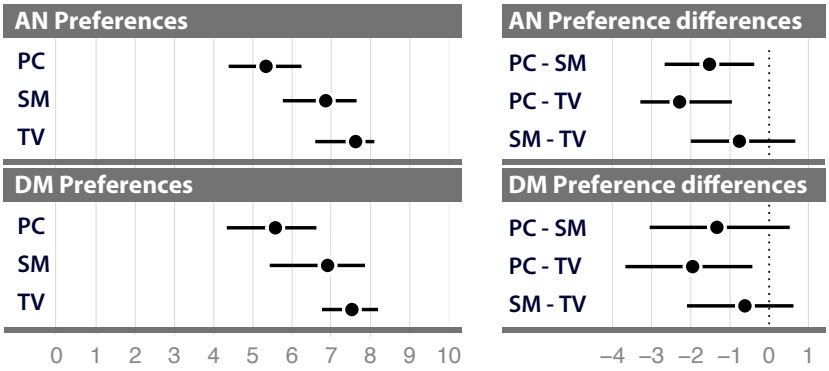


Figure 3.11: *Left:* Mean rating for each technique, for the analytic (AN) and for the decision (DM) tasks. *Right:* Mean differences in ratings between each pair of techniques — a positive value indicates a preference for the technique on the left. Error bars are 95% CIs (n=21).

technique on the right (conversely, a negative value indicates a preference for the right technique). All CIs are 95% BCa bootstrap confidence intervals.

Q3a. For analytic tasks, PC appears as the least preferred (mean preference [0-10]: PC 5.3, SM 6.9, TV 7.6). Our data does not provide enough support for a difference between SM and TV.

Q3b. For decision making, results suggest that participants prefer TV over PC (mean preference [0-10]: PC 5.6, SM 6.9, TV 7.5). We do no have enough evidence to draw other conclusions.

3.4.1.4 Q4 – Subjective Choice Assessment

Figure 3.12 reports how participants evaluated the choice they made in the decision-making task, depending on the technique used. Each horizontal panel presents the results according to a different choice assessment metric (see Section 3.3.10.2). On the difference plots, a positive value indicates a higher average rating for the technique on the left. All CIs are again 95% BCa bootstrap confidence intervals.

Q4a. There is no evidence of a major difference between techniques in terms of average participants' satisfaction with their choice (mean satisfaction [0-10]: PC 7.6, SM 7.8, TV 7.2). We cannot conclude as to the direction of the effects, but the differences are likely no more than ± 1 point on an 11-point Likert scale.

Q4b. The data is also inconclusive regarding participants' confidence in their choice (mean confidence [0-10]: PC 7.4, SM 7.5, TV 7.7), except we know that large effects are implausible (likely not above ± 1.5 points).

Q4c. Concerning perceived easiness, the precision of our estimates is again low, but it is not implausible that decisions made with TV are perceived as easier to make on average than with PC and SM (mean easiness [0-10]: PC 6.6, SM 6.9, TV 7.5). However, the evidence is rather weak.

Q4d. The data suggests that on average, participants may be slightly more attached to their choice

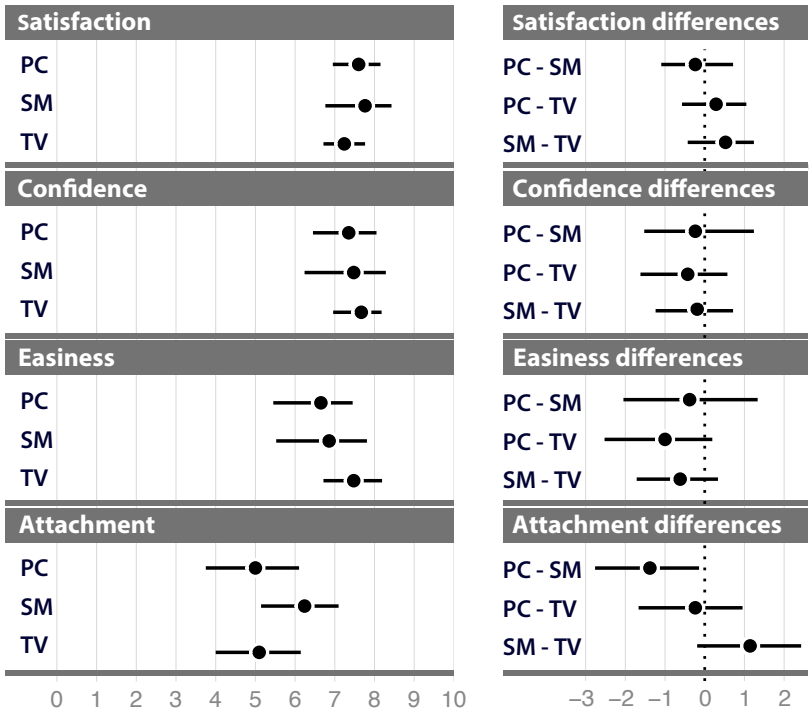


Figure 3.12: *Left:* Mean rating for each choice assessment metric and each technique. *Right:* Mean differences between each pair of techniques — a positive value indicates a benefit (e.g., higher choice satisfaction) for the technique on the left. All error bars are 95% CIs (n=21).

if they made it using SM than if they used either PC or TV (mean attachment [0-10]: PC 5, SM 6.2, TV 5.1). There is no evidence for a major difference between PC and TV in terms of attachment.

3.4.2 Additional analyses

We now report additional (unplanned) analyses, to better understand in what respects the three techniques differ.

3.4.2.1 Time-on-Task for Decision Making

When framing our research questions, we reasoned that time-on-task was of secondary concern for decision making, as the answers themselves seemed more important than the time it took to reach them. Time-on-task is also difficult to interpret for open-ended tasks, as increased time can be a sign of both increased difficulty and increased engagement.

However, the three techniques turned out to be hard to distinguish in terms of decision accuracy and subjective choice assessment. The effects there are likely small (i.e., likely not more than a $\pm 10\%$ difference in accuracy and $\pm 15\%$ for subjective metrics, see Figures 4.6 and 3.12), requiring a large

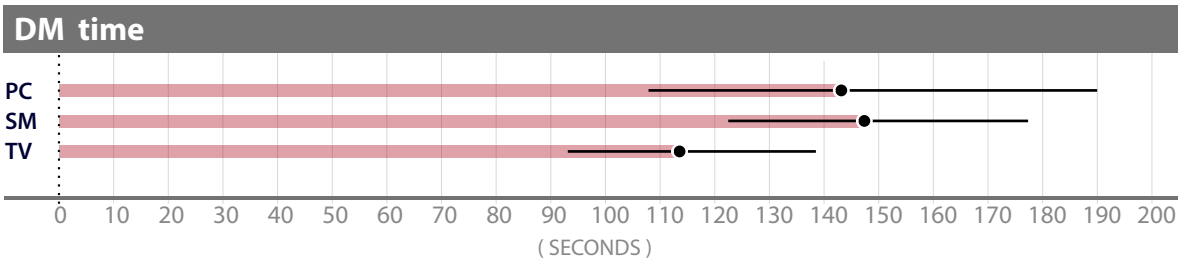


Figure 3.13: Average time (in sec) spent on the decision making task for techniques PC, SM and TV. All error bars are 95% CIs (n=21).

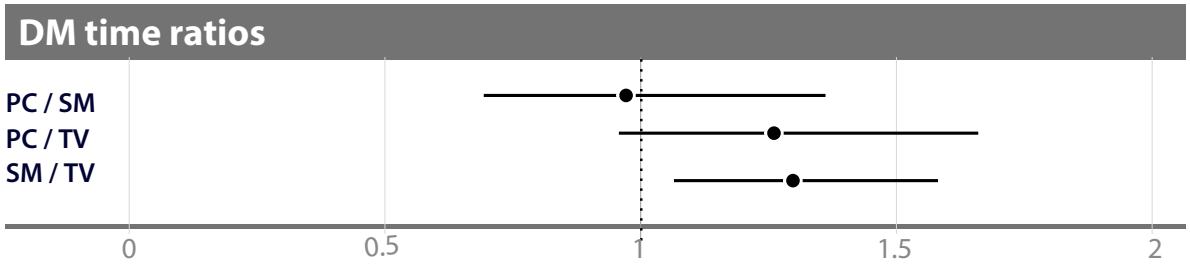


Figure 3.14: Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21).

statistical power to be estimated reliably. Therefore, the time metric can be a useful differentiating factor. Time can also be of particular interest when decisions have to be made rapidly.

Figures 3.13 and 3.14 show the average amount of time it took participants to make their choice with each technique. The analysis method was the same as for the analytic tasks (subsubsection 3.4.1.2). As we can see, there is some evidence that decisions were made more rapidly with TV than with the other two techniques: both SM and PC took on average 1.3 times longer, with 95% CI [1.1, 1.6] for SM and 95% CI [0.96, 1.7] for PC (mean times in sec: PC 140, SM 150, TV 110).

3.4.2.2 Differences in preferences

We further wanted to explore whether people changed their preference for the technique when switching from analytic to decision tasks. Figure 3.15 shows both mean and absolute difference in preference for each technique. The bottom plot results show that participants changed their preference for each technique 2 Likert scale points on average when performed the decision making task. The upper plot results are inconclusive on whether this change was positively or negatively towards the decision task.

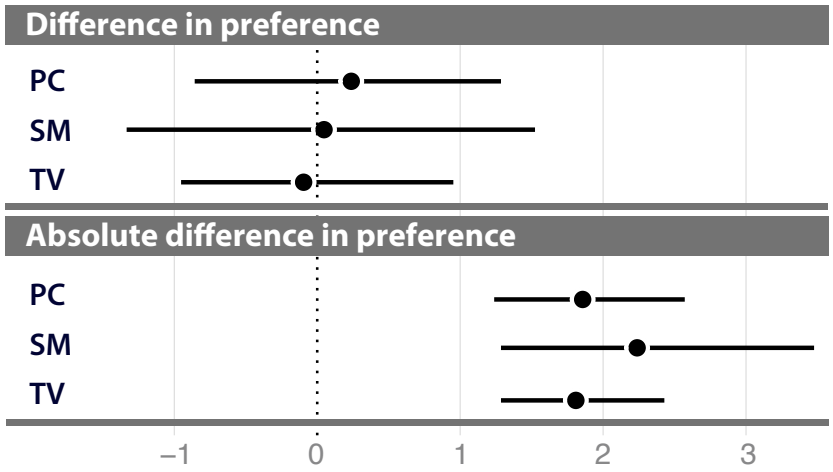


Figure 3.15: *Upper:* Mean difference in preference for techniques PC, SM and TV from analytic to decision tasks — positive values indicate that people increased their preference for the technique when used it for a decision task. *Bottom:* Mean absolute difference in preference for a technique PC, SM and TV from analytic to decision tasks — positive values indicate that people changed their preference for the technique when used it for a decision task. All error bars are 95% CIs (n=21).

3.4.2.3 Deviation of preferences

We also wanted to explore the amount of variation in participant preferences.

Analytic tasks: The standard deviation of preferences for PC was σ = 2.2, 95% CI [1.8, 2.7], for SM & VA: σ = 2.2, 95% CI [1.7, 2.9] and for TV & VA: σ = 1.7, 95% CI [0.96, 2.8].

Decision tasks: The standard deviation of preferences for PC was σ = 2.7, 95% CI [2.3, 3.1], for SM & DM: σ = 2.8, 95% CI [2.1, 3.8] and for TV & DM: σ = 1.7, 95% CI [1.3, 2.2].

The results indicate that in most cases there was high variability across individuals, but people seem to rate TV more consistently in decision tasks, than in analytic tasks.

3.4.2.4 Qualitative feedback

A text field allowed participants to optionally provide justifications for their technique ratings on decision tasks, and 13 out of the 21 participants did so. Two raters (co-authors) independently segmented text responses into comments about a particular technique and classified these comments into positive or negative (Cohen's kappa = 0.90 for segmenting+classification). The 13 respondents produced a total of 50 comments. We review these comments here.

PC received *9 negative* comments, characterizing PC as hard to use for comparing alternatives, as well as for searching, isolating and selecting a single alternative. PC received *4 positive* comments, on how polylines allow performing a quick evaluation of individual alternatives.

SM received *5 negative* comments, mainly on the complexity of the visual representation, and on the amount of information presented that can be overwhelming. SM received *11 positive* comments

referring to its advantages in overview tasks making it possible to see patterns and trade-offs between attributes, as well as its ability to filter multiple attributes at the same time.

TV received *4 negative* comments, mainly because alternatives can only be sorted by one attribute, making it hard to perform overview comparisons that take into account all attributes. TV received *17 positive* comments, stating that it was easy and straightforward in a range of elementary tasks, e.g., comparing and identifying alternatives, or isolating and selecting them. One participant also found TV's support for manual reordering of alternatives very useful for making decisions.

Overall, PC received the largest number of negative comments, mostly because it did not allow to easily compare alternatives. SM and TV received the largest number of positive comments, mostly because they supported well overview (for SM) and elementary tasks (for TV). Meanwhile, some comments about SM appeared strongly negative (e.g., "extremely difficult" to understand at first), while none of the negative comments on TV seemed to have reported strong drawbacks. TV has also received the largest number of positive comments (17 vs. 11 and 4). Although no strong conclusion can be derived from this observation alone, it appears consistent with the preference ratings (Figure 3.11).

3.5 Discussion

To verify that participants understood and used effectively the visualizations, we first evaluated the visualizations on analytic tasks. All techniques yielded close-to-perfect accuracy, indicating participants were able to use them effectively. There were however large differences in completion times: SM was slowest for value retrieval and range tasks but by far the fastest in correlation tasks. The lower performance of SM in the two low-level analytic tasks can be explained by the lower resolution of SM's axes compared to PC (see Figure 3.7), and by the difficulty of dealing with two axes concurrently. As one participant noted "*I felt I had to pay more attention to which axis corresponded with which variable, and my eyes were on the axes while dragging on the dots*". On the other hand, the efficiency of SM for correlation tasks is not surprising, as scatterplots are known to convey correlation effectively [271, 303]. Also, SM shows all pairwise correlations simultaneously, while both PC and TV required manual attribute reordering to examine them in sequence. Though PC is considered competent in conveying correlations [210, 224, 303], it was outperformed by TV both on time and accuracy.

The second part of the evaluation involved decision making tasks. Overall, we found our techniques to be comparable across metrics, with a slight advantage for TV. Participants also preferred TV over PC overall. Participants reported being more attached to choices they made with SM on average, a result that needs to be confirmed by further studies. The reasons for this are currently unclear, although one explanation could be that SM's better support for overview tasks (confirmed by our results with the correlation task) made participants more confident that they did not miss a particularly interesting alternative. However, this difference was not captured by the confidence metric.

The result that participants made their choice faster on average with TV than with SM or PC, is not anticipated by the analytic tasks. One possibility is that they felt more engaged with SM and

PC, thus spending more time with these techniques. However, it seems unlikely that they lingered on the tasks out of interest for the datasets or techniques, since the decision task involved the same techniques and the same type of dataset as the analytic tasks, and the task was administered three times in a row. It seems more plausible that TV simply allowed participants to reach answers faster.

The decision metrics overall showed a larger variability in responses compared to the analytic tasks. Ideally when evaluating visualizations using *multi-attribute choice tasks*, the results should clearly indicate which visualization technique is better to support such tasks. However, even though the results showed a slight advantage for TV, decision metrics (such as accuracy, time, or technique preference) had larger variability in responses than their equivalent analytic ones. One, perhaps obvious, explanation for this difference between analytic and decision tasks is that *multi-attribute choice tasks*, especially when they involve personal preferences, are inherently subjective. For example, unique personality traits of a participant asked to identify a correlation or to retrieve a value are less likely to influence a response than when asked to choose her ideal holiday package. Thus, it is possible that preference-based decision metrics are not sensitive enough to capture small but practically meaningful differences across conditions. In addition, participants may not be able to perfectly express (or be aware of) their criteria preferences, which likely adds further noise to our accuracy metrics. The insensitivity of the metrics used in combination with the small sample size of the laboratory study make it harder to capture differences across conditions that likely exist [106].

3.6 Conclusion

This chapter investigated how to rigorously evaluate multidimensional visualizations for their ability to support multi-attribute choice tasks. It first identified which of the existing multidimensional visualizations are compatible with such tasks, and set out to evaluate three elementary visualizations: parallel coordinates, scatterplot matrices and tabular visualizations. The three visualizations seemed to be comparable on most metrics, but tabular visualizations allowed participants to reach decisions faster and, overall, to be a compelling choice, despite the low attention they have received in the literature on multidimensional visualization.

The proposed evaluation method consists of first giving participants low-level analytic tasks, to ensure that they properly understood the visualizations and their interactions. Then multi-attribute choice tasks are examined, through multiple objective and subjective metrics, including a decision accuracy metric based on the consistency between the participant's choice and their self-reported preferences for attributes. Moreover, although decision time is typically not central in assessing decision support, the study results show that it can be used as a tie-breaker when visualizations achieve similar decision accuracy. The study results also suggest that indirect methods (attachment) for assessing choice confidence may allow to better distinguish between visualizations compared to direct ones (rate confidence). Limitations of the methods and directions for future work were finally discussed, such as the need for more sensitive metrics of decision support.

The evaluation of Parallel Coordinates, Scatterplot Matrices and Tabular Visualizations was added to the table of decision support evaluations in visualizations (Appendix B, discussed in section 2.4). As it is indicated by the three ■ marks next to each visualization, this chapter addressed all identified limitations in the evaluation of previous works. This chapter used generic, dataset-independent visualizations, sensible visualization baselines, actual decision tasks, and metrics that examine decision accuracy.

The use of multi-attribute choice tasks provided additional insights not necessarily anticipated by analytic tasks, e.g., speed advantage for TV, or attachment in choice with SM. Nevertheless, multi-attribute choice tasks also gave responses of greater variability and lesser accuracy than analytic tasks. To understand multi-attribute choice tasks at a deeper level, the next chapter will attempt to make a direct comparison of accuracy between analytic and multi-attribute choice tasks.

To capture possible differences between analytic and multi-attribute choice tasks, it is necessary to eliminate some factors that make multi-attribute choice tasks harder to analyze: subjectivity and sample size issues. An important source of subjectivity was the fact that the task relied on personal preferences (e.g., choosing holiday package). Such tasks may give the impression it is not necessary to give an optimal and carefully examined response. Next chapters will focus on less subjective multi-attribute choice tasks, where personal preferences play a smaller role. Second, to achieve larger statistical power, similar to most experiments in decision research, the setup will be on a crowdsourcing platform. Such platforms are a relatively new and promising way of accessing a larger number of participants that allows effective evaluations of visualization tools [63, 65, 221, 278, 332]. Also, to make multi-attribute choice tasks easier to analyze the number of attributes will be limited to two. As explained in section 2.1.3, to de-emphasize the focus on a large number of attributes, such multi-attribute choice tasks will now be called, simply, *choice tasks*.

Chapter 4

The Role of Instructions in Decision Making Tasks

Given differences in accuracy found in Chapter 3, this chapter focuses more on comparing response accuracy between analytic and choice tasks. The analytic and choice tasks examined in Chapter 3 had different types of instructions. The analytic instructions were short and rather abstract, e.g., *"Which two attributes have the strongest positive correlation? "*, whereas the choice task was presented using a narrative framing *"Imagine now that you want to go for vacations and you are searching for holiday packages. Holiday packages can have different attributes such as the overall price, the quality of the hotel you will stay, the nightlife (number of restaurants, clubs) or how unique is the landscape [...]* (followed by a description of all 8 attributes) *... Now you need to choose the ideal holiday package for you according to your preferences on the attributes we discussed [...] "*. Comparing these two instructions raises an interesting question: What could be the effect of the narrative itself on participant responses? On the one hand, it is possible that a narrative can engage a participant to perform more accurately, feeling that this is a real decision task for their future vacations. On the other hand, it is also possible that longer instructions can confuse or distract a participant from the actual task. Besides, a more salient description of any task may introduce additional subjectivity and thus noise in participant responses.

Considering that narratives could have played a role in participants responses, one solution could be to investigate choice tasks alone without using narrative instructions. However, it is rather unclear how to do so. For example, asking directly "Choose a house." to participants who have no real intention to buy a house may not trigger the right reaction, and lead instead to follow-up questions such as "for what purpose", "for whom", "with what criteria". In fact, it seems that all previous visualization studies that examined choice tasks used some form of narrative instructions: some were rich and long, e.g., instructing participants to choose a nursing home for the 75-year-old injured father of their best friend [524]; and others were simpler, immersing participants in shopping scenarios [51], or in

managerial decisions [35, 127]. Narratives appear to be an essential and rather inevitable component when providing instructions for choice tasks. However, it seems that no study in visualization research examined the effect of narratives used in experiment instructions on participants' performance.

A primary goal in this chapter will be to investigate whether narrative instructions could be a reason why decision tasks induce noisy and less accurate responses. Since narratives seem to be inherently connected to decision instructions, narratives will also be added in analytic task instructions, to investigate if it is the use of narratives in general, or the decision making scenario that affects performance. For example, the abstract correlation in the grades dataset (used in the analytic evaluation in Chapter 3) would be now framed as: "Imagine that you are a teacher that wants to understand whether your students' performance in Math is related to their performance in Physics.". If narrative analytic instructions appear to trigger responses of similar (or lesser) accuracy to narrative decision instructions, it would be an indicator that narratives could have played an important role in the difference between choice and analytic tasks observed in Chapter 3.

A secondary goal in this chapter will be to investigate which is the best way of presenting a task to maximize accuracy. It is generally known that the way of framing a task can have important implications on how people respond [475], but there is a lack of previous work in the visualization literature on how to frame instructions for both analytic and decision tasks. In particular, several instructions will be compared: *abstract instructions* (e.g., find the data point with the minimum X value); *semantic instructions* with minimal information about the dataset (e.g., find the cheapest available house); and *narrative instructions*. Enriching a task with context can be more engaging and make the task easier to understand, but it is also possible that longer instructions can be more demanding regarding time and patience and make the goal of the task harder to grasp. Nevertheless, given the positive effects of narratives often mentioned in the literature [202, 203, 263], it is expected that the possible advantages of a narrative (e.g., engagement) will outweigh the limitations (e.g., clarity).

Most parts of the following sections were previously published in [135]. Thus any use of "we" in this chapter refers to Evanthia Dimara, Anastasia Bezerianos and Pierre Dragicevic.

4.1 Narratives in Information Visualization

There is evidence that humans can more easily make sense of the world through narratives, i.e., coherent sequences of events [202, 203, 263]. Researchers and practitioners have already started to use narratives in the context of data analysis and communication, in order to improve data understanding and engagement of the users [242, 413]. But narratives can also be used during visualization evaluation, in the form of a backstory in the task instructions, to help simulate the real use of a system and elicit a more representative user behavior. We next discuss studies on question-wording and the use of narratives in information visualization.



Figure 4.1: Storytelling outdoors, Concord Library, 1970 by Paul Ife Horne, appeared in Concord.

4.1.1 Question wording

Psychologists have long been interested in the effects of question-wording. Some of the work in this area has focused on how question framing can affect reasoning and judgment, for example in terms of causal attribution [507]. Other work has focused on how to best design surveys to get reliable responses. Past work suggests that conciseness and context are both desirable [346], but it remains unclear how to strike the right balance between the two. In addition, guidelines for survey design may not directly translate to information visualization evaluation. For example, issues like desirability bias (i.e., respondents trying to give socially acceptable answers) are key in survey design [266] but likely less relevant to visualization evaluation.

4.1.2 Visualization narratives

A currently popular line of research in information visualization suggests that complementing interactive visualizations with stories about data (visualization narratives) can turn data exploration into a more engaging and educational experience [242, 244, 413, 421, 486]. For example, narratives can be used

for explaining changes in complex temporal networks [39], or for promoting user engagement during data exploration [66]. Two particularly compelling application areas are journalism [64] and science communication [313]. A number of tools have been proposed to help authors design visualization narratives and interleave textual stories with visual elements [223, 243, 296, 413].

Our work significantly differs from visualization narratives in that we explore the use of narratives during the *evaluation* of visualizations, not during their actual use. Thus our "end users" are study participants, not data consumers. Our narratives invite users to put themselves in a hypothetical situation (e.g., being a real-estate analyst, or a house buyer), which is not the case in typical visualization narratives. Finally, although the narratives we explore provide context about the datasets, they do not refer to trends and patterns in the data itself.

4.1.3 Illustrative use cases

Visualization designers and researchers have long used narratives in the form of illustrative use cases in order to convey a tool's functionalities in a way that is more accessible and persuasive than a factual description. In the research literature, we can find two types of narratives used in this context: *analytic narratives* and *decision-making narratives*. Typical analytic narratives involve an expert who seeks to understand a dataset within a domain like cyber security [163] or business priority analysis [93]. For example, an ocean forecaster may want to analyze the Red Sea dataset for glider path-planning [233]. Alternatively, non-specialists can be involved, such as a person who seeks to grasp their nutrition habits [201]. Common decision making narratives involve a person seeking a house to buy [513], a prospective student choosing a university [93, 201], or a company executive choosing the location of a new factory branch [33].

Compared to the narratives we study, illustrative use cases have in common the hypothetical situation component, but target different end users (article readers) and significantly differ in content (fictional data exploration activities).

4.1.4 Narratives in visualization evaluation

Narratives are sometimes used in information visualization evaluation, for example in instructions to briefly present participants with a choice task. For instance, Yi et al. [525] asked their participants to choose a cereal brand, using the same narrative as in their illustrative case study. Other minimalist forms of narratives that are solely used to attribute a meaning to the datasets, are also common. For example, to evaluate HomeFinder, Williamson et al. [513] used questions such as "what neighborhood has the most expensive houses?".

Full narratives are commonly used in evaluations of decision-support visualization systems. For example, in order to evaluate a tool for software release plans, Aseniero et al. [35] asked participants to take the role of a project manager and choose the optimal plan. In order to evaluate a tool for preferential choices, Bautista and Carenini [51], immersed participants in shopping scenarios involving

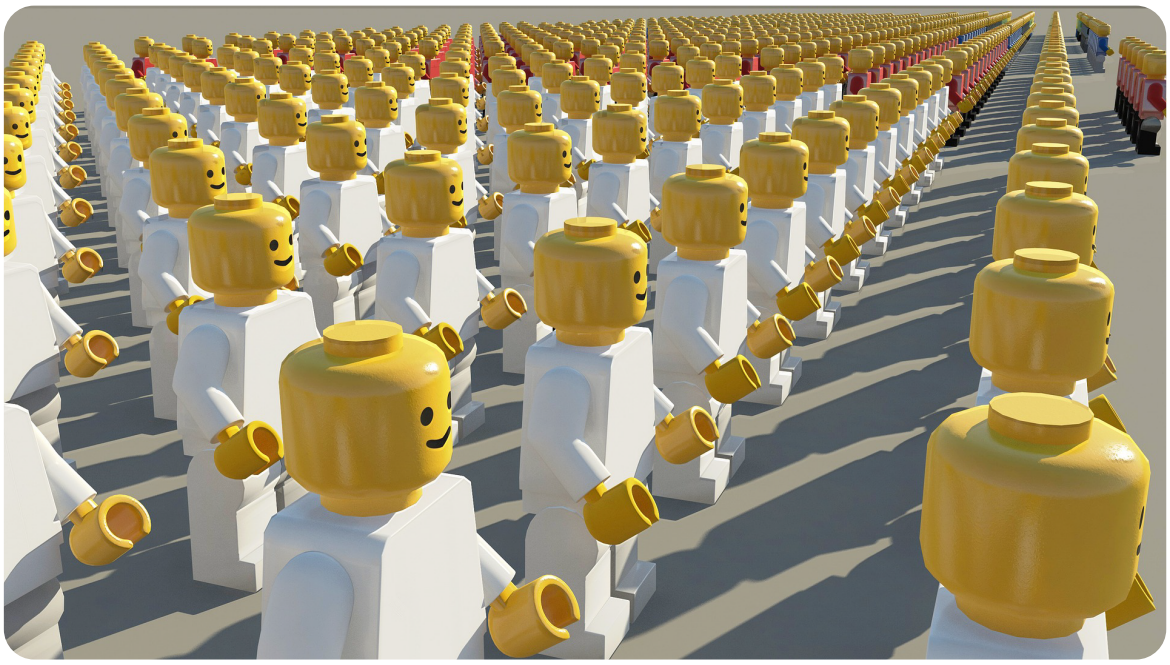


Figure 4.2: Crowdsourcing. source:https://pixabay.com/en/crowd-lego-staff-choice-selector-1699137

television sets, houses or cell phones, and put them in a situation of finding a hotel to stay in Vancouver. Similarly, Daradkeh et al. [127] asked participants to make hypothetical investments.

Although information visualization researchers sometimes use narratives in their evaluations, we do not know the effect of decision making narratives, or, more generally, how other narratives affect performance or evaluation, both in lab settings and crowdsourcing.

4.2 Narratives in crowdsourcing

Crowdsourcing platforms are a promising way of accessing a large and diverse pool of participants, allowing rapid evaluations of visualizations [63, 65, 221, 278, 332]. However, engaging crowdworkers and obtaining high quality responses can be challenging [150, 245]. In particular, task instructions in a remote study where the instructor has no way of helping or motivating the participants, should be designed with extra care. We thus need to better understand how task instructions affect the quality of responses in the evaluation of visualization tools.

There are two main reasons why narratives could be used in crowdsourcing task instructions. One reason is that that narratives presumably help simulate a "natural context" [507] and thus, a more representative use of the system, which is especially important when evaluating domain-specific and decision-support visualization systems as seen before. As we discussed in the introduction, if we want to carry out a crowdsourced evaluation of a system meant to help customers choose a car, plain instructions such as “select the best car” may not put crowdworkers in the right frame of mind.

Therefore, it may seem more suitable to use a decision-making question framing and provide a narrative context that could help participants simulate a hypothetical purchase situation, or recall a similar situation from the past.

Another reason is that a narrative can possibly provide benefits in terms of enhanced motivation, attention and engagement, even if the evaluation's aim is to investigate how a generic visualization tool supports basic analytic tasks. These benefits could possibly translate into improved task comprehension and higher-quality responses. Improving the quality of responses is especially important in crowdsourced studies. Although crowdourcing is now widely accepted as an evaluation platform [221, 278, 366], the overall quality of responses can be low, which either leaves investigators with poor data to analyze or forces them to reject a large proportion of responses [245].

A number of strategies have been suggested to improve the quality of responses in crowdsourced studies. A common approach consists of only recruiting contributors with high reputation, possibly subjecting them to qualification tests [221], and using verification questions to detect lack of diligence [221, 278, 362, 366]. Optimal payment strategies have also been explored [57, 235, 319], but studies suggest that higher monetary rewards increase the quantity but not the quality of responses [57, 319]. Other recommendations include using short task durations [150, 278] while avoiding breaking down tasks into meaningless chunks [61, 319]; paying attention to experiment design [283]; and providing sufficiently challenging, personalized and easy to understand tasks [245]. Even though much of previous work has emphasized the importance of providing clear, meaningful and engaging tasks, to our knowledge there is no study investigating whether the use of task narratives in visualization evaluation, and in decision support evaluation in particular, can yield measurable benefits.

4.3 Experiment

Our goal was to explore the effect of adding task context, in particular in the form of narratives, in a crowdsourced visualization evaluation. To identify what effects stem from adding minimal context as opposed to more complex backstories, we compared: *i)* providing no context whatsoever about the data, *ii)* providing minimal semantic context on the data (e.g., referring to houses rather than abstract data points), and *iii)* adding backstory narratives that also justify the purpose of the task. We used two types of narratives, as well as an additional control condition that will be explained later on.

Participants were assigned to one of the five context conditions and performed three basic visualization tasks using scatterplot visualizations. To assess the merits of the different context conditions, we used objective performance metrics that measured participants' ability to perform and understand the tasks, as well as subjective metrics based on self-reported impressions.

4.3.1 Dataset and visualization

Our study involved simple datasets with two quantitative dimensions. The datasets were small-sized (21 data points each) artificial datasets created manually using spreadsheet software.

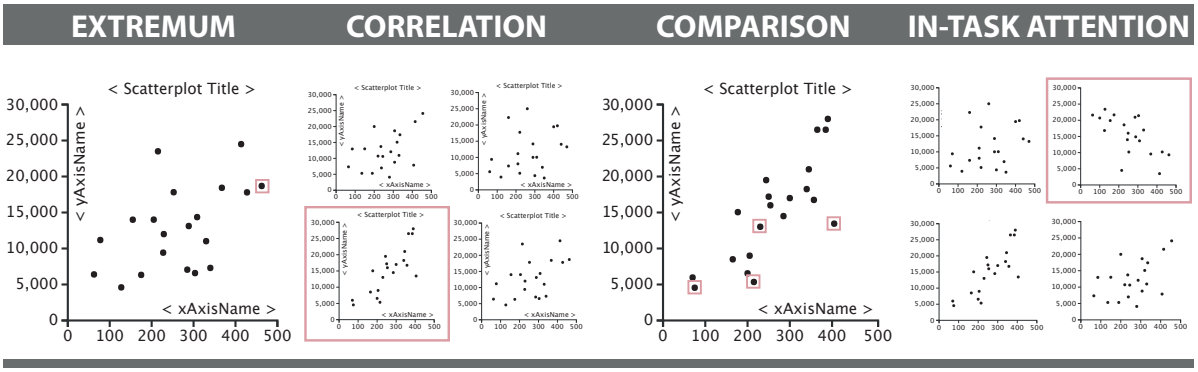


Figure 4.3: Stimuli used in each task (Ext, Cor and Com), and in the in-task attention test. Correct answers are annotated in color. Axes were labeled (X, Y) for ABS, and ($size\ m^2$, $price$ (\$)) in all other context conditions. The title was $Diagram\ Z : Datapoints$ in ABS, and was $Diagram\ Z : Houses$ in SEM (all tasks) and DM-NAR (Ext, Cor tasks). In all other conditions the title was $Agency\ Z : Houses$. Z was an integer (1, 2, 3, or 4) identifying the scatterplot.

For the experimental stimuli, we used a 2D scatterplot visualization, as it is a standard information visualization technique for presenting multiple data points along two dimensions [347]. Moreover, our study involves overview tasks (discussed in section 4.3.2), and, according to our findings in chapter 3, scatterplots are particularly effective in such tasks. The scatterplots supported basic interactions that depended on the task and will be described in the next subsection.

4.3.2 Tasks

We used three basic visualization tasks adapted from taxonomies of low-level information retrieval task [20, 402, 504]:

- An **Extremum** task (**Ext**), where participants had to find the data point with highest value according to the X dimension (see the leftmost scatterplot in Figure 4.3).
- A **Correlation** task (**Cor**), where participants had to find the scatterplot with the highest correlation among four different ones (see the second panel in Figure 4.3).
- A **Comparison** task (**Com**), where participants had to compare data points across their two dimensions simultaneously (see the third panel in Figure 4.3). The task consisted of finding a data point without any "competitor", a competitor being defined as a data point that has both larger X and smaller Y. The task had four possible correct answers. The comparison task is particularly important, since it will be framed as a choice task

In one of the experiment conditions (described in section 4.3.3), all these low-level analytic tasks will be adapted using a decision making backstory narrative. The most important task is the comparison which will be framed as a choice task.

As said before, the scatterplots supported basic interactions. In the Ext and Com tasks, hovering over a data point highlights it in light gray, displays horizontal and vertical projection lines, and

overlays the data point's X and Y values on the axes. Participants gave their answer by clicking on a data point, after which its color changed to green. For the Cor task, scatterplots were highlighted in light gray when hovered. Participants selected their answer by clicking on one of the four plots, after which the selected plot changed to green. In all tasks, participants could either confirm their choice by clicking on a "next" button, or change their selection.

4.3.3 Context conditions

As context for our datasets, we decided to use scenarios involving the real estate market. Our choice is based on both the nature of the house price/house size tradeoff that is easy to understand, and its use in previous evaluations of both analytical and decision making visualization systems [51, 513].

For each task, instructions were split into two pages on the Web form: **page 1** displayed introductory and background information relevant to the task; and **page 2** showed the task question and the visualization. Participants were allowed to navigate back-and-forth between the two pages.

Each task came in five different variants, one per context condition. Scatterplots were identical or had minor label differences (see caption of Figure 4.3), while the major differences were in the text instructions on page 1 and page 2. Table 4.1 shows the complete text instructions for all of the context conditions except O-NAR, covered later on. The study employed:

- A **Decision making narrative** (**DM-NAR**), where page 1 contains a narrative that asks participants to put themselves in the situation of a house buyer and, given some criteria and constraints, to make choices. Questions were of the type *"Given what you read, which house would you buy?"*.
- An **Analytic narrative** (**AN-NAR**) condition, where page 1 contains a narrative that asks participants to put themselves in the situation of a real estate analyst and to find answers to analytical questions. An example of a question on page 2 would be *"Given what you read, which house would be the most attractive to your customers?"*.
- A **Simple Semantics** (**SEM**) condition, where the data points were houses and dimensions were price and size. Questions were of the type *"Which is the biggest house?"*.
- An **Abstract** (**ABS**) condition, where the dataset has no specific meaning. Both page 1 and page 2 use abstract wordings. An example is *"Which is the data point with the largest value of X?"*.

In the two narrative conditions AN-NAR and DM-NAR, participants had to read the narrative on page 1 to be able to interpret the question on page 2. Thus, participants who do not read the text on page 1 carefully enough will see their performance negatively impacted. In the SEM condition, in contrast, the question is self-contained for tasks Ext and Cor (but not Com, see Table 4.1). Because the narrative conditions differ from SEM in two respects (the presence of a narrative, and the necessity to read page 1 to be able to answer any question), we introduced a fifth, intermediary condition:

- **Optional Narrative** (O-NAR) condition, a hybrid control condition where page 1 is identical to AN-NAR and page 2 is identical to SEM. An example of a question would be *"Which is the biggest house?"*. Thus, despite the presence of a narrative on page 1, participants did not need to read it to answer the task question.

In order to rule out poorly framed narratives, we tested and refined them through crowdsourced and in-person pilot studies.

4.3.4 Objective performance metrics

In this section, we describe how we measured performance. All metrics were devised before data was collected.

4.3.4.1 Accuracy

For all tasks we used a normalized measure of *accuracy* ranging from 0 to 1. We preferred quantitative to binary metrics because of their higher statistical power. We assigned 1 to correct answers (Figure 4.3, in color). For other answers, we gave a score depending to how close they are to the right answer.

In the Ext task, where participants needed to find the data point with the largest X, each of the 21 data points got a score of $S\ (\frac{X-X_{min}}{X_{max}-X_{min}})^2$, where X is x-coordinate of the chosen data point, X_{min} is the minimum x-coordinate of the plot, and X_{max} is the x-coordinate of the correct answer.

In the Cor task, where participants needed to identify the highest correlation, we assigned a score of $S\ \frac{C-C_{min}}{C_{max}-C_{min}}$, where C stands for the correlation of the selected plot, C_{min} is the lowest correlation and C_{max} if the correlation of the correct plot. This time we did not square the score because incorrect correlations were much lower than the correct one.

In the Com task, where users needed to identify a data point without competitors, i.e., a non-dominated point (see dominance definition in chapter 2.2.1), we assigned a score of $S\ (\frac{D_{max}-D}{D_{max}-D_{min}})^2$, where D stands for the number of points that dominate the selected point, D_{max} is the maximum number of points that dominate any point in the dataset, and D_{min} is the minimum number (zero in our case).

4.3.4.2 In-task attention

Since a lack of diligence or a poor understanding from participants may not always translate into incorrect responses, we used attention as a secondary measure of performance. As a proxy for in-task attention, we measured participants' ability to recall the options presented to them in the correlation task (see the rightmost panel in Figure 4.3). The test was administered after all tasks were completed. We asked participants to identify which plot was not presented to them before. As we can see in Figure 4.3, the correct answer has a negative correlation, whereas all options presented previously had a positive correlation. The in-task attention measure is likely linked to other factors such as task comprehension. We expect that participants who understood the correlation relation would be able to recall its visual pattern and identify the "unexpected" negative relation.

Since all incorrect answers were about equally wrong, for the in-task attention metric we assigned a binary score of 1 for the correct answer and 0 for all other answers.

4.3.4.3 Post-Task attention

Because researchers may want to conduct longer experiments than ours and because narratives may yield participant fatigue (or alternatively, abstract tasks could cause a loss of interest), we also measured participant's attention after the tasks. We administered at the end of the experiment an independent instructional manipulation check where participants needed to read instructions very carefully to get a correct answer [362]. As before, we assigned a binary score of 1 for the correct answer and 0 for all other answers.

4.3.4.4 Metrics not considered

We did not consider task completion time as a metric, as it would be difficult to interpret in the context of our study. This is because, depending on the context condition, longer task completion times could be either an indication of lower performance (e.g., in the ABS condition), or an indication of higher motivation and engagement (e.g., in the DM-NAR condition).

4.3.5 Subjective metrics

We used subjective metrics of confidence, easiness, enjoyability, and usefulness as a complement to the previous metrics. All responses were reported on a 7-point Likert item.

- *Confidence:* After each task, we asked participants to report their confidence in their answer.
- *Easiness:* We also asked them to rate the perceived difficulty of each task. Since we wanted all scores to reflect a positive direction, we referred to easiness rather than difficulty.
- *Enjoyability:* After all tasks had been completed, we asked participants to report how much they enjoyed the job overall.
- *Usefulness:* We asked participants to report to what extent they thought the diagrams would be useful if they wanted to buy a product. The goal was to examine if a richer context makes tasks more meaningful and change participants' perspective on the utility of the visualization tested.

4.3.6 Experiment design

The experiment followed a mixed design. The independent between-subjects variable was context (ABS, SEM, O-NAR, AN-NAR, DM-NAR). The independent within-subjects variable was the task (Ext, Cor, Com). Each participant performed all three tasks in the same order: Ext, Cor and finally Com, accounting for what we thought was an increasing level of difficulty. Since each participant was assigned to a unique context condition, they each saw the three tasks with the same type of context provided.

4.3.6.1 Procedure

We ran the experiment as a Crowdflower job[1]. Participants opened an external 12-page Web form. They first performed the Ext task, consisting of two pages as previously mentioned. They selected their answer as described previously. On the following page, they rated their confidence and task easiness. They followed the same process for the Cor and Com tasks. Once they finished the 3 tasks, participants rated the enjoyability of the job, the usefulness of the diagram, and were given the instructional manipulation check on the same page. On the next page, they were given the in-task attention test. On the last page, they provided basic demographic information, and were finally given a completion code to paste in crowdflower to complete their job. Participants spent on average 7 minutes on the job and were given a reward of 60 US cents.

4.3.6.2 Crowdsourcing quality control

Although a common crowdsourcing practice is to reject jobs from participants whose performance is abnormally poor or who failed attention tests, we accepted and analyzed all jobs[2]. The reasons are twofold: *i)* since different conditions are expected to yield different levels of attention and performance, excluding low-quality jobs would bias our results; *ii)* we seek to improve the overall quality of *all* submitted jobs, with the hope that fewer jobs will need to be rejected in the future.

4.3.6.3 Participants

Our total sample consisted of 405 highly rated (level 3) Crowdflower contributors. Sample size per condition ranged from n=80 to n=83 (for a planned sample size of n=80). Figure 4.4 summarizes participants' self-reported demographics.

4.3.7 Research questions

Prior to data collection, we framed four research questions and hypotheses. Since our hypotheses were not derived from a theory, we refer to them as "expectations" [140].

Q1 *Is a decision making narrative better than an analytical narrative? This was our main research question. The purpose was to compare analytic and decision narratives regarding participant response accuracy. As explained in the introduction, narratives are inherently connected to decision tasks, so the only way of comparing the two is by adding narratives in the analytic tasks. Although we did not expect large differences for the Ext and Cor tasks, we expected that* DM-NAR *would outperform* AN-NAR *for the Com task, since this task involves mental operations (dominance recognition) typical of everyday decision making tasks.*

[1] `https://www.crowdflower.com`

[2] Three jobs, however, had to be rejected because their duration went over the 30-min limit imposed by the crowdsourcing platform.

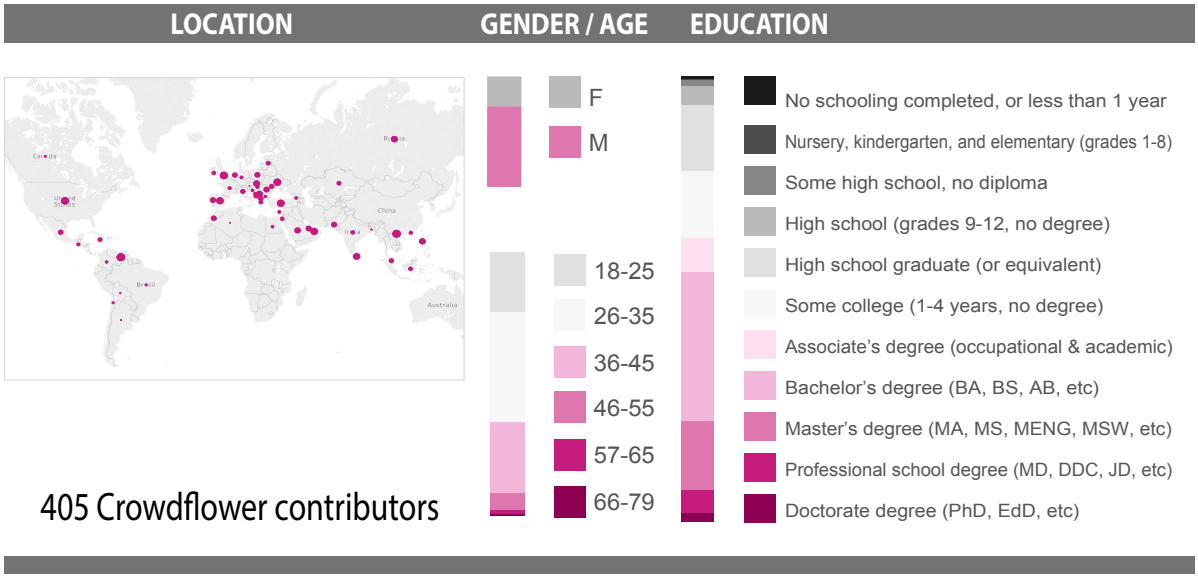


Figure 4.4: Self-reported demographics of our participants.

Q2 *Does adding a narrative (on top of minimal semantics) help? This was our secondary research question.* We expected that the benefits of narratives (e.g., higher engagement) would outweigh their costs (e.g., higher attention demands).

Q3 *Does adding minimal semantics help?* Assuming we find an effect of narrative, the purpose was to determine how much of the effect is simply due to the fact that the narrative assigns a meaning to the dataset and its dimensions. We expected benefits when adding minimal semantics alone.

Q4 *Should the question refer to the narrative?* The purpose was to better understand the reason behind any effect of narrative we may find. For example, if narratives happen to yield poorer performance but the control condition O-NAR does not, it could mean that the problem comes from participants not reading the narratives. In addition, if O-NAR alone yields improvements, it could mean that task-irrelevant narratives are sufficient to motivate participants.

4.3.8 Overview of results

As in the previous chapter, we analyze, report and interpret all our inferential statistics using interval estimation [140]. Experimental stimuli, data, and analyses are available at `http://www.aviz.fr/narratives`.

Before we turn to our main research questions, we first give an overview of all our results. We report the sample mean for each condition according to our objective performance metrics (accuracy, in-task attention, and post-task attention) and our subjective metrics (confidence, perceived easiness, overall job enjoyability and perceived usefulness of the visualization). In addition to sample means, we report 95% confidence intervals (CIs) indicating the range of plausible values for the population mean [140]. See Figure 4.5 for help on how to interpret overlaps in CIs. For in-task and post-task

Condition	Task	Page 1	Page 2
ABS	Ext	You will be asked to answer a few questions about data. In the next page you will see many data points displayed in a diagram.	Which is the data point with the largest value of X?
	Cor	Now you will see four diagrams with data points. You will be asked to compare them.	In which of these four diagrams is Y most related to X?
	Com	Now you will see one of the previous diagrams again. You will be asked a question that requires identifying "competitors". In our case, a data point is a competitor of another data point if it has both larger X and smaller Y.	Select a data point that has no competitor.
SEM	Ext	You will be asked to answer a few questions about houses. In the next page you will see many houses displayed in a diagram.	Which is the biggest house?
	Cor	Now you will see four diagrams with houses. You will be asked to compare them.	In which of these four diagrams is price most related to size?
	Com	Now you will see one of the previous diagrams again. You will be asked a question that requires identifying "competitors". A house is a competitor of another house if it is both bigger and cheaper.	Select a house that has no competitor.
AN-NAR	Ext	You will be asked to answer a few questions about houses. Imagine that you are a real estate analyst and you need to understand the house market. You focus on extremely rich customers who seek to buy a house that is as big as possible. In the next page you will see the houses currently on the market, displayed in a diagram.	Given what you read, which house would be the most attractive to your customers?
	Cor	Now you want to focus on regular customers who are not necessarily very rich. You want to investigate how reliable some real estate agencies are. You will see four diagrams with houses. Each diagram shows the houses proposed by a different agency. An agency that sets arbitrary prices is NOT reliable. While in a reliable agency, price is very related to size.	Given what you read, which of these four real estate agencies is the most reliable?
	Com	Now you will see one of the previous diagrams again. It shows the houses offered by the best agency. You need to report on their best deals. A good deal is a house that has no "competitor". A house is a competitor of another house if it is both bigger and cheaper.	Given what you read, select a house that is a good deal.
DM-NAR	Ext	You will be asked to make a few decisions about houses. Imagine you are moving to a new city and you need to buy a house. You are extremely rich and you want your house to be as big as possible. In the next page you will see the houses currently on the market, displayed in a diagram.	Given what you read, which house would you buy?
	Cor	You don't have as much money as you initially thought. So before buying a house, you need to find a reliable real estate agency. You will see four diagrams with houses. Each diagram shows the houses proposed by a different agency. An agency that sets arbitrary prices is NOT reliable. While in a reliable agency, price is very related to size.	Given what you read, which of these four real estate agencies would you choose?
	Com	Now you will see one of the previous diagrams again. It shows the houses offered by the best agency. You will finally get to choose your house. A good choice is a house that has no "competitor". A house is a competitor of another house if it is both bigger and cheaper.	Given what you read, which house would you buy?

Table 4.1: The instructions text in each experiment condition. The condition O-NAR has identical page 1 with AN-NAR and identical page 2 with SEM condition.

attention, we use Wilson's confidence intervals for a single proportion. For all other metrics, we use BCa bootstrap confidence intervals.

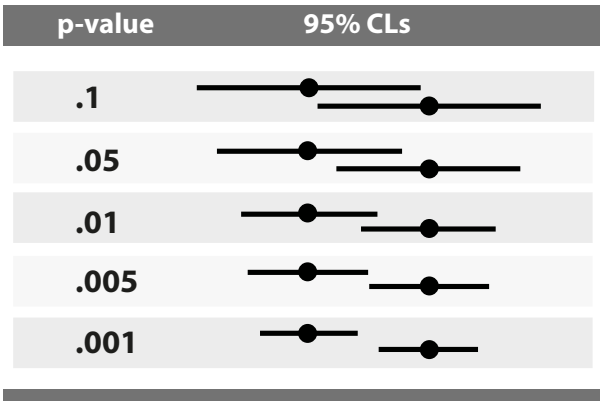


Figure 4.5: Indicative chart showing the correspondence between degree of CI overlap and *p*-values for independent samples (after [290]).

4.3.8.1 Accuracy

Mean accuracy scores are shown in Figure 4.6, with tasks on columns and conditions on rows. The first column shows scores averaged across all three tasks. As we can see on this column, crowdsourced participants were fairly accurate overall (scores of 0.7–0.8 out of 1). However, it appears that participants who were given the DM-NAR narrative performed less accurately on average than those who were only given minimal context (SEM) or no context at all (ABS). The other narrative AN-NAR may have also performed worse than SEM, but the evidence is much weaker.

Regarding the extremum (Ext) task, both narrative conditions AN-NAR and DM-NAR appear less accurate on average than all other conditions (Ext mean accuracy: ABS 88%, SEM 86%, AN-NAR 77%, DM-NAR 75%, O-NAR 86%). For the correlation (Cor) task, AN-NAR appears worse than SEM, while DM-NAR and O-NAR may also be worse than SEM, but the evidence is weaker (Cor mean accuracy: ABS 63%, SEM 67%, AN-NAR 55%, DM-NAR 58%, O-NAR 59%). For the comparison (Com) task, DM-NAR is clearly worse than AN-NAR. For this task, AN-NAR appears to outperform ABS (Com mean accuracy: ABS 82%, SEM 86%, AN-NAR 92%, DM-NAR 80%, O-NAR 88%).

4.3.8.2 In-task attention

As we can see in Figure 4.7, participants exhibited a better recall of the correlation task when given minimal semantics (SEM) than no context (ABS), suggesting they were paying more attention. However, adding a narrative (AN-NAR or DM-NAR) to the semantics decreased their recall. The decrease is less evident but possible for the control condition O-NAR where the narrative was not required to perform the task (mean in-task attention: ABS 66%, SEM 84%, AN-NAR 66%, DM-NAR 65%, O-NAR 75%).

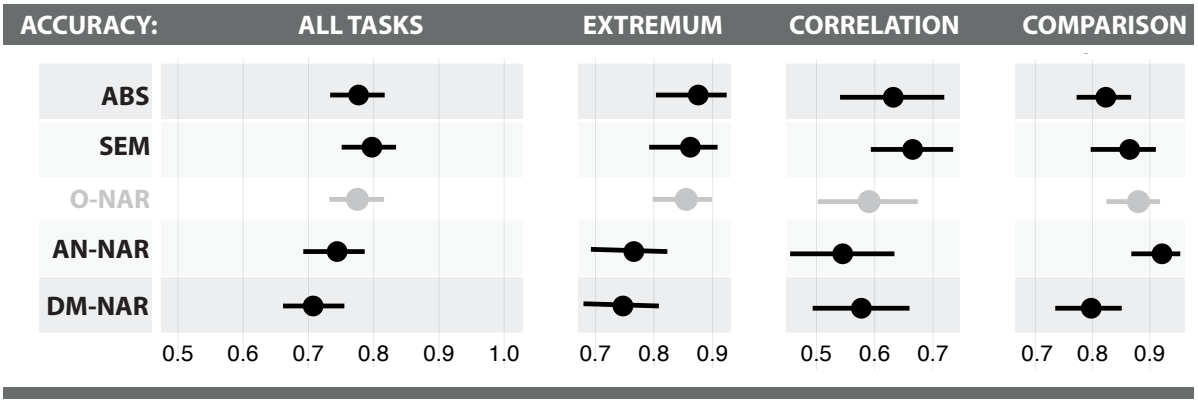


Figure 4.6: Accuracy per task and condition. Error bars are 95% CIs.

4.3.8.3 Post-task attention

As we can see in Figure 4.7, the results are mostly inconclusive regarding post-task attention. There is, however, some weak evidence that adding a narrative when it is not needed (O-NAR) may make people less attentive after they perform the task compared to providing only minimal semantic context (SEM) (mean post-task attention: ABS 60%, SEM 67%, AN-NAR 65%, DM-NAR 63%, O-NAR 56%).

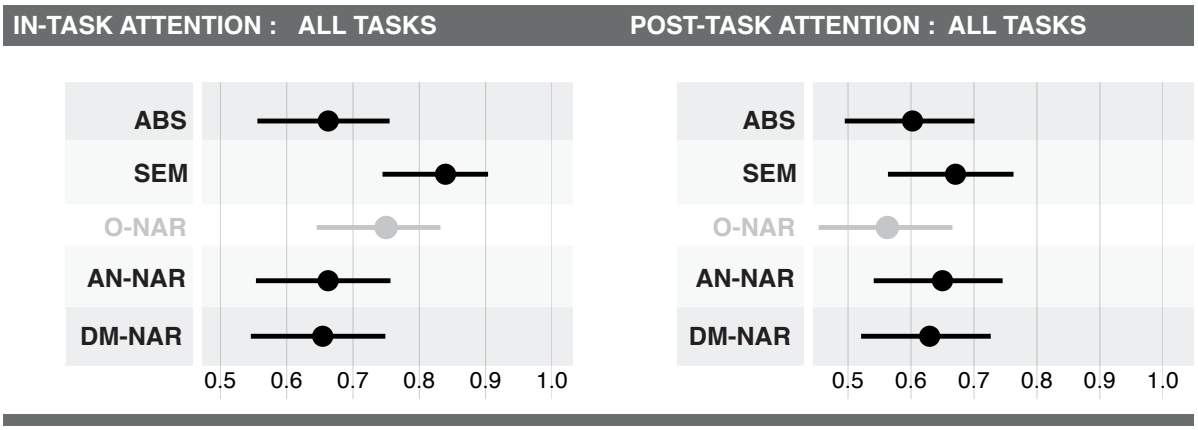


Figure 4.7: In-task and post-task attention. Error bars are 95% CIs.

4.3.8.4 Confidence

Figure 4.8 reports confidence scores normalized between 0 and 1. Confidence was overall high (0.7–0.8), but participants were on average less confident when provided no context (ABS). We did not observe this differences for the Ext task (Ext mean confidence: ABS 87%, SEM 89%, AN-NAR 84%, DM-NAR 85%, O-NAR 83%), but it is clear for Cor (Cor mean confidence: ABS 66%, SEM 78%, AN-NAR 74%, DM-NAR 75%, O-NAR 76%) and remarkably large for Com, with the remaining conditions yielding comparable confidence scores (Com mean confidence: ABS 58%, SEM 78%, AN-NAR 80%, DM-NAR 80%, O-NAR 78%).

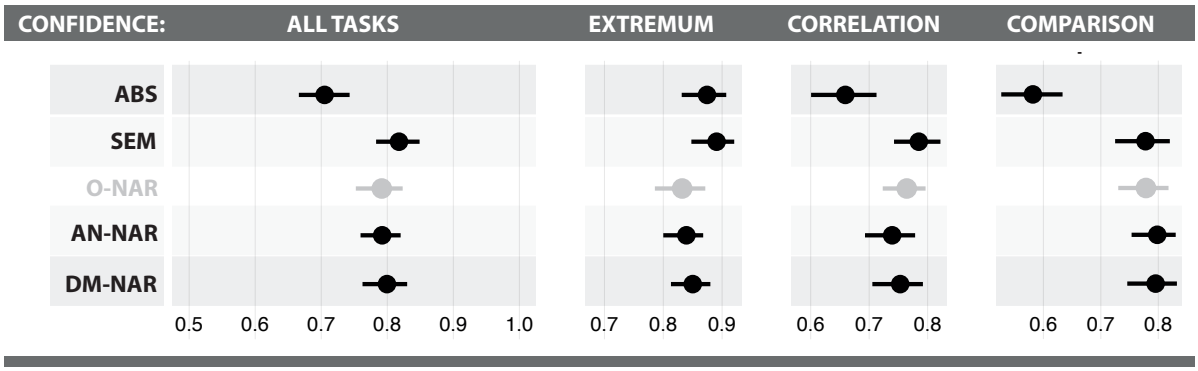


Figure 4.8: Reported confidence scores. Error bars are 95% CIs.

4.3.8.5 Easiness

Figure 4.9 provides some evidence that without context (ABS) the tasks appear harder overall. Although there is no visible difference for the Ext task (Ext mean easiness: ABS 77%, SEM 77%, AN-NAR 72%, DM-NAR 79%, O-NAR 70%), for the Com task the difference is clear (Com mean easiness: ABS 61%, SEM 72%, AN-NAR 74%, DM-NAR 74%, O-NAR 70%). There is also some evidence that participants found the control condition O-NAR a bit harder overall, especially for the Ext and Cor tasks (Cor mean easiness: ABS 65%, SEM 70%, AN-NAR 69%, DM-NAR 68%, O-NAR 62%). Finally, for the Ext task, the use of a DM-NAR narrative may have made the task appear easier compared to the use of a AN-NAR narrative.

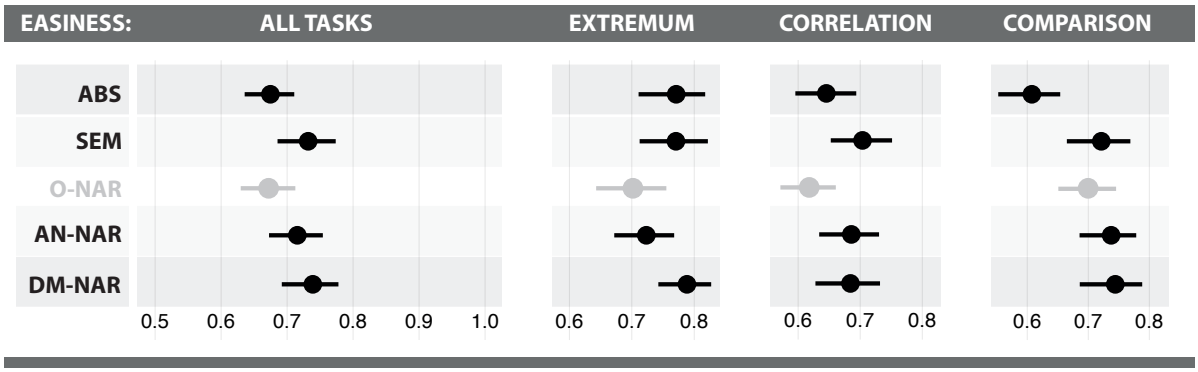


Figure 4.9: Reported easiness scores. Error bars are 95% CIs.

4.3.8.6 Enjoyability and usefulness

Figure 4.10 provides good evidence that when no context is provided on the visualization tasks (ABS), participants find the overall job less enjoyable (mean enjoyability: ABS 74%, SEM 83%, AN-NAR 84%, DM-NAR 83%, O-NAR 82%) and rate the visualization as less useful (mean usefulness: ABS 69%, SEM 82%, AN-NAR 78%, DM-NAR 80%, O-NAR 78%) than when any type of context is provided.

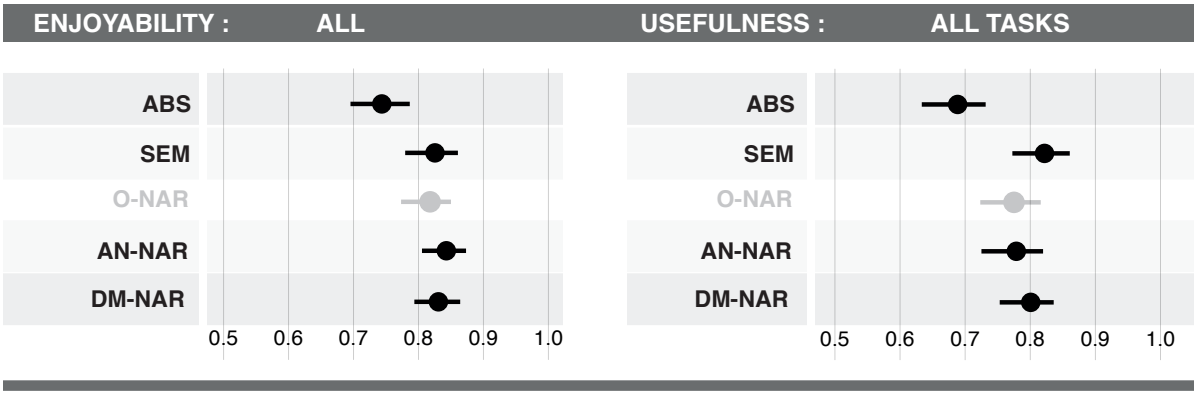


Figure 4.10: Enjoyability and Usefulness

4.3.9 Planned analyses

In the previous section we gave an overview of all our results and identified several patterns, but due to the many comparisons involved some of these patterns may not be reliable. In this section, we report on more focused comparisons based on our previously stated research questions. All the analyses in this section were planned before data was collected.

As before, we report sample statistics with 95% CIs. For dichotomous variables (in-task and post-task attention), we report proportion differences and compute CIs using score intervals for difference of proportions and independent samples. For continuous variables (all other metrics), we report differences in means and BCa bootstrap confidence intervals.

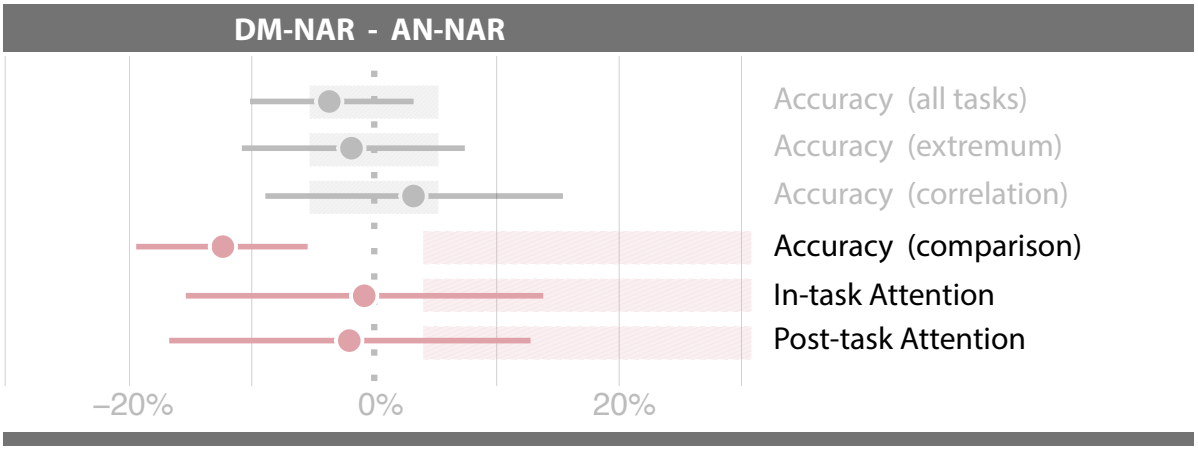


Figure 4.11: *In gray:* Mean differences in accuracy between DM-NAR and AN-NAR across tasks and for Ext and Cor; *In color:* Mean differences in accuracy, in-task and post-task attention between DM-NAR and AN-NAR. Positive values indicate a benefit for DM-NAR. Error bars are 95%CIs.

4.3.9.1 Q1: Is a decision-making framing better than an analytic one?

To answer this question, we compared AN-NAR and DM-NAR in terms of their performance on the comparison (Com) task. We previously justified our focus on the Com task by explaining that we did not expect large differences between the two narratives for Ext and Cor. For context, Figure 4.11 shows, in gray, the differences across all tasks and for the tasks Ext and Cor. The results are consistent with our conjecture, although there is a larger uncertainty concerning the correlation task.

The rest of the figure (in color) shows the differences for the Com task, a task that mimics a real decision task and for which we expected the decision making narrative (DM-NAR) to outperform the analytic narrative (AN-NAR). But contrary our expectations, participants performed remarkably worse when given a DM-NAR narrative. Concerning in-task or post-task attention, our results are inconclusive.

4.3.9.2 Q2: Does adding a narrative help overall?

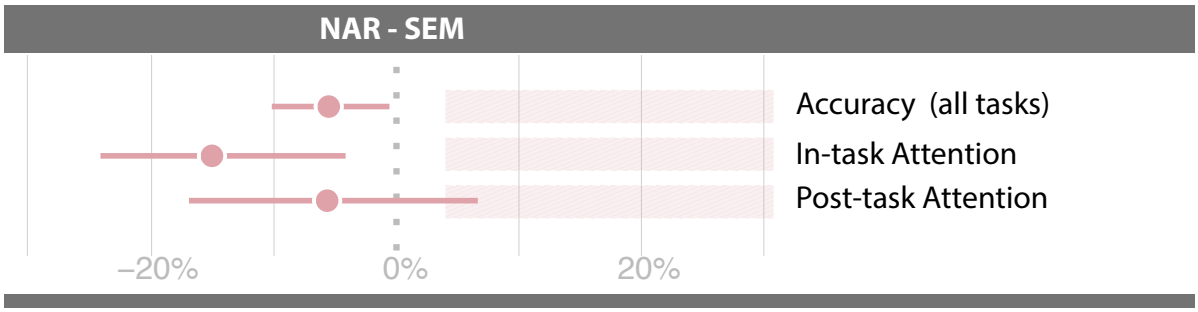


Figure 4.12: Mean differences in accuracy, in-task attention and post-task attention between all NAR conditions combined and SEM. Positive values indicate a benefit for narratives. Error bars are 95%CIs.

To answer this question, we performed a contrast between the (SEM) condition and all narrative conditions (O-NAR, AN-NAR, DM-NAR) combined. Here too, we expected a positive effect of narrative across all metrics. However, the results in Figure 4.12 go contrary to our expectations. Adding a narrative on top of dataset semantics makes participants less accurate. It also makes them less able to recall the correlation task, suggesting lower in-task attention. We do not have enough data to conclude that narratives also reduce post-task attention, but this remains a possibility.

4.3.9.3 Q3: Does adding minimal semantics help?

To answer this question, we compared the ABS and SEM conditions for each objective performance metric: accuracy (averaged across all tasks), in-task attention and post-task attention. The results are shown in Figure 4.13. The shaded areas indicate our initial expectations, i.e., a positive effect of SEM on all metrics. Contrary to our expectations, we found no evidence that adding semantics has a noticeable effect on participants' accuracy. We also found no evidence of a strictly positive effect on post-task attention, although the uncertainty on this metric is rather large. Nevertheless, participants

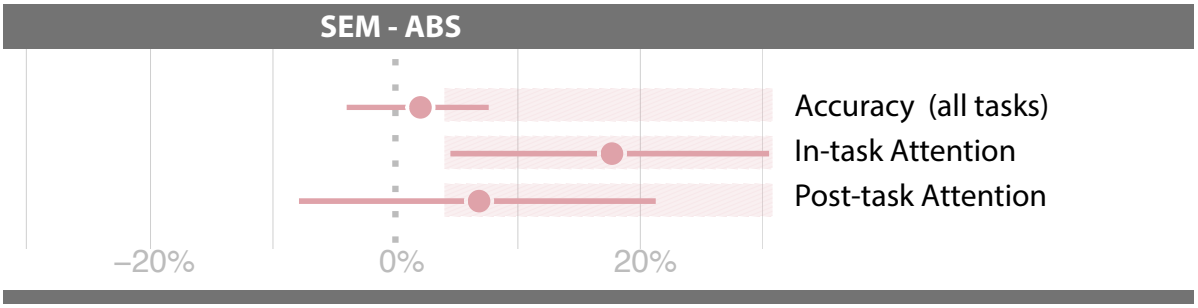


Figure 4.13: Mean differences in accuracy, in-task attention and post-task attention between SEM and ABS. Positive values indicate a benefit for SEM. Error bars are 95%CIs.

better recalled the correlation task, suggesting that adding minimal semantic context can have positive effects on in-task attention.

4.3.9.4 Q4: Should the question refer to the narrative?

To answer this question, we compared O-NAR and AN-NAR. We expected that participants would be more accurate when reading the narrative is not required to carry out the task (O-NAR). The results in Figure 4.14, do not indicate a clear direction for an effect, and only suggest that the difference is rather small. We also thought participants would pay less attention when the narrative is not required. The data is mostly inconclusive. There is only very weak evidence that this could have been the case for post-task attention, but that the opposite pattern may have occured for in-task attention.

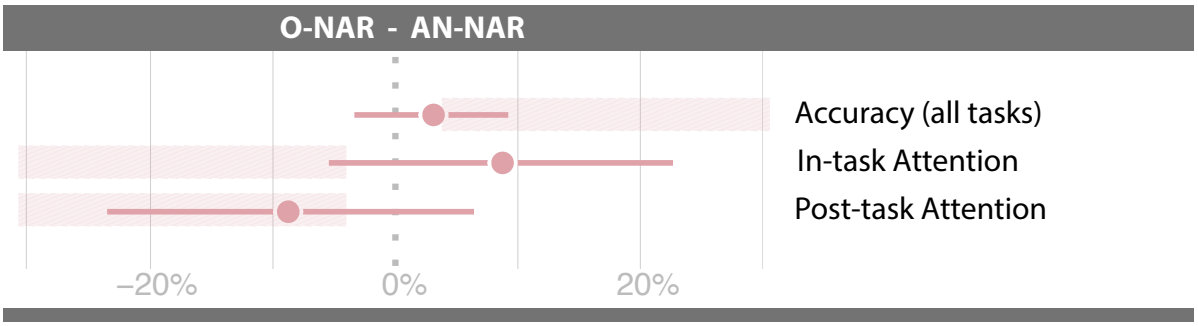


Figure 4.14: Mean differences in accuracy, in-task attention and post-task attention between O-NAR and AN-NAR. Positive values indicate a benefit for O-NAR. Error bars are 95%CIs.

4.4 Discussion

This section discusses two topics, the general use of narratives in experiment instructions and the decision making narratives in particular.

4.4.1 On the use of narratives in experiment instructions

In our study participants were fairly accurate overall, with average scores of 0.7–0.8 out of 1 across all tasks and narrative conditions (see Figure 4.6). Subjective task easiness scores were between 0.6–0.8 out of 1. Thus it seems that the difficulty of the tasks was properly calibrated overall.

Our findings suggest that providing task context in the form of data semantics or narratives does not necessarily improve the overall quality of responses: regarding accuracy, data semantics do not seem to help much, and the narratives we used can even harm.

Our experiment only allows us to speculate about the reasons for these findings. First, as we discussed before, crowdworkers generally appreciate succinct instructions [150]. An otherwise simple task can appear more demanding in attention and time if it requires reading a long (in crowdsourcing standards) piece of text beforehand. Second, experienced contributors are generally used to performing abstract and mechanical tasks since these abound on crowdsourcing platforms. The fairly good performances we observed for abstract conditions do suggest contributors were overall able to understand the context-less tasks and willing to carry them out.

It is unclear whether contributors simply skipped the narratives. On the one hand, results of our post-task attention test confirm that not all our contributors read all instructions carefully (only 50–70% passed the test, see Figure 4.7). On the other hand, we did not find evidence that this was the reason for the lower task accuracy (research question **Q3**, Figure 4.14). However, Figure 4.6 does suggest that for the Ext task, asking the question in a way that does not require reading the narrative can help.

Despite these results, we have strong evidence that adding data semantics improves subjective experience on a range of metrics (confidence, perceived easiness, enjoyability, and perceived usefulness of the visualization). Adding data semantics also seems to help crowdsource contributors pay more attention and possibly better understand the tasks. Although this was not reflected in our accuracy scores, it remains possible that a difference would be detected with more statistical power. Thus providing minimal semantic information about datasets and their attributes can make tasks more salient, engaging and meaningful.

However, our backstory narratives did not yield measurable subjective benefits compared to data semantics alone. Thus, even though crowdsourcing contributors appreciate working with meaningful data, they may not be particularly interested in more elaborate narratives and may prefer to focus on carrying out their task.

Overall, our study provides compelling reasons for incorporating data semantics in crowdsourced evaluations of visualizations, i.e., stating what the datasets and their dimensions mean. But until further studies are carried out to nuance or contradict our findings, it seems safer to use elaborate narratives parsimoniously, unless there are clear reasons to do so. Such reasons include the evaluation of domain-specific and decision-support visualization systems.

4.4.2 On decision making narratives

In the comparison task, which decision making instructions framed it as a *choice task*, participants had to identify the best (dominant) house offer. All participants were explicitly instructed that a "good" house offer is the one for which there is no alternative house in the dataset that is both bigger and cheaper. All participants were crowdsourcing workers also explicitly instructed that their payment will depend on whether their answers are correct or not. The house offers were presented as black dots in a 2-D scatterplot with dimensions “price” and “size” and no other information available that could further influence their (e.g., house photo, location, etc.). These participants were divided into two groups: analytic and decision group. The only difference between the two groups was the instruction framing. In the analytic group, participants were asked to imagine they were hypothetical real estate agents who needed to identify the house that would most likely please their clients. In the decision group, participants were asked to imagine they were hypothetical house buyers who needed to choose the best house to buy. The decision group performed significantly less accurately than the analytic group.

A question that arises is: why a visualization task framed as a decision problem makes people perform less accurately than when the same task is framed as an analytic one? One could argue that a decision framing allows subjective preferences to influence the response. This could be the case in a real house choice task. For example, people could indeed prefer smaller houses. However, it seems unlikely that in the context of an artificial crowdsourcing choice task (with houses presented as black dots and where payment depends on correctness) that participants were inaccurate due to their true preference for smaller houses. Another possible answer could be that maybe the narrative of a real-estate agent was more motivational than the narrative of a house buyer. However, these two narratives were also tested in different tasks than a choice task. Participants performed with similar accuracy when identifying a correlation in the house dataset, regardless of being instructed to be house agents or house buyers.

Therefore, this work showed compelling evidence that choice tasks can be more error-prone than equivalent analytic tasks. Also, although narratives can generally be a reason for less accurate responses, narrative instructions can not justify this difference, since the narrative analytic instructions gave more accurate responses.

4.5 Conclusion

This chapter explored the effects of providing task context when evaluating visualization tools using crowdsourcing. Crowdworkers were given i) abstract information visualization tasks without any context, ii) tasks with added semantics to the dataset, and iii) tasks with two types of backstory narratives: an analytic narrative and a decision-making narrative. Contrary to the stated expectations, there was no evidence that adding data semantics increases accuracy, and further, backstory narratives can even decrease accuracy. Adding dataset semantics can, however, increase attention and provide

subjective benefits in terms of confidence, perceived easiness, task enjoyability and perceived usefulness of the visualization. Nevertheless, narratives did not appear to provide additional subjective benefits.

In addition, although the presence of a narrative can in principle be a reason for less accurate responses, it seems that it is not the main reason why decision tasks were so prone to errors. Narratives were also added in the analytic framing, but participants were more accurate than when the visualization task was framed as a decision problem.

The experiments in Chapters 3 and 4 cannot sufficiently explain why decisions appear to induce more errors. One reason could be that the decision accuracy metrics used in the previous experiments are not very informative. The *multi-attribute choice tasks* studied in chapter 3 (holiday packages) and in this chapter (houses) were evaluated with metrics that rely on the notion of *dominance* (as defined in chapter 2.2.1). These tasks involved a dataset that consists of superior (or dominant) and inferior (or dominated) alternatives. Participants either "succeeded" by choosing a dominant alternative or "failed" by choosing an inferior one. When participants "fail", a simple explanation could be that they did not identify the dominance relation, thus, they did not understand the visualization. However, according to the findings of the narratives experiment, this explanation is rather implausible, since participants successfully identified the same dominance relationship in the analytic framing. So, if the visualization does not seem to be the problem, what makes participants "fail"? The dominance-based metrics used to evaluate decision quality are probably not enough to help us answer this question. The next chapter will explore alternative methods of measuring decision quality.

CHAPTER 5

Detecting Cognitive Biases in Visualization Systems

The chapter 4 contributed a finding that has implications for the evaluation of visualizations for decision making: choice tasks appear to be more error-prone than equivalent analytic tasks. Nevertheless, evaluating decisions using preference-based metrics (chapter 3) or dominance metrics (chapter 4) gives limited information of the source of these errors.

Generally, a visualization is considered effective if it helps people extract accurate information [84, 530] and thus help to make an informed decision. However, as we discussed in the section 2.3,

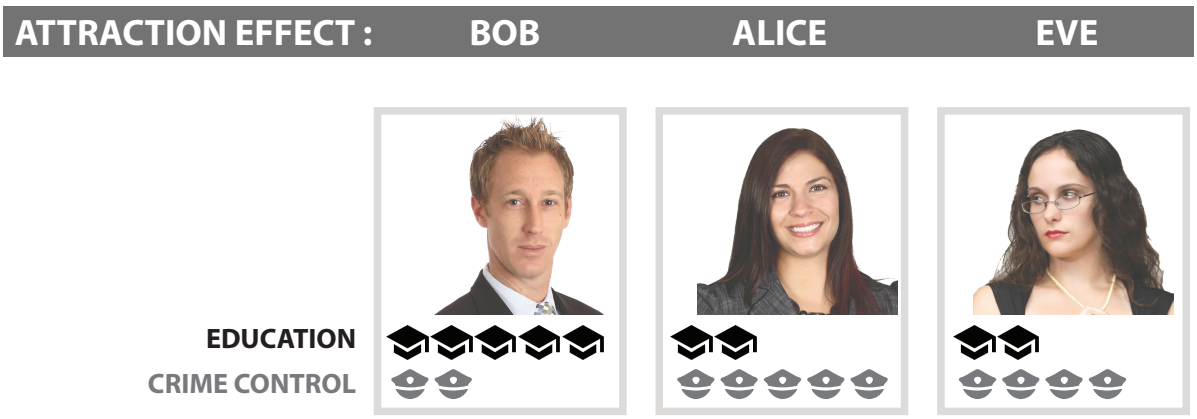


Figure 5.1: Example of an attraction effect in elections: Bob has an excellent education plan, while Alice is very strong in crime control. The addition of Eve, a candidate similar but slightly inferior to Alice, raises Alice's attractiveness as a candidate. This irrelevant option is called a *decoy*. (Photos Benjamin Miller, FSP Standard License, icons by Ivan Boyco, CC-BY license)

full access to information does not necessarily yield good decisions [263]. It seems that when people are facing a decision puzzle, they tend to resort to heuristics, i.e., *"simple procedures that help find adequate, though often imperfect, answers to difficult questions"* [263]. While heuristics can be very effective [190], they can also lead to cognitive biases [263]. Therefore, in order to fully understand how information visualizations can support decision making, it is important to study how they interact with cognitive biases. This chapter investigates an example of such a bias.

Suppose citizens are voting for primary elections and need to choose between candidates Bob and Alice (Figure 5.1). Bob has a solid education plan, but not much concern for crime control. In contrast, Alice's education plan is weak but she has an excellent strategy for crime control. If both education and safety are important to them, this can be a difficult choice. Now suppose there is a third candidate, Eve. Like Alice, Eve focuses more on crime control than education, but her crime control strategy is not as good as Alice's. O'Curry and Pitts [358] used a similar choice task in a study, and showed that adding Eve as an option shifted participants' preference towards Alice instead of Bob. This shift in preference called the *attraction effect* (also known as the *decoy effect* and the *asymmetric dominance effect*), is a cognitive bias whereby people tend to favor the option for which there exists a similar, but slightly inferior, alternative. Like other cognitive biases, the attraction effect leads to irrational decisions and has important implications in many areas such as politics and advertising.

This chapter focuses on the attraction effect for three reasons. First, the attraction belonds to the **biases of a faulty choice task** category (see ■ #CHOI in Appendix A and section 2.3.5). This category of biases, although not yet studied in visualization research, involves systematic errors in people's choices. Consequently, detecting choice biases directly affects visualization systems that target decision support. Second, it is one of the most studied cognitive biases in fields such as psychology, consumer research and behavioral economics. Third, although the attraction effect has been extensively studied, these studies generally employ very small sets of alternatives (typically three) and text formats, so it is still unknown whether the bias generalizes to data visualizations. Although some visual representations have been considered, there is conflicting evidence and a heated debate on whether the effect generalizes [174, 241, 434, 523]. Some argue that the effect occurs only in numerical stimuli [174], e.g., when attributes are presented in tables. Whereas others argue that it is generic and robust, and can be observed in many contexts such as visual judgments in shapes [467], oral instructions [422], or even among animals when they choose their food [295]. This debate suggests that the attraction effect is far from being fully understood and needs to be investigated from a variety of perspectives.

This chapter investigates whether people can be subject to a cognitive bias while using a visualization tool. In particular, it will focus on whether the attraction effect has implications for information visualization design. In the previous example, voters' decision is influenced by the presence of Eve, which is inferior in all respects and therefore irrelevant to the choice. If, in the same way, someone uses a visualization to choose among several options (e.g., when buying an apartment [513]), will the presence of inferior choices affect their decision? In other words, does the attraction effect transfer to visualizations?

Most parts of the following sections were previously published in [134] and [133] . Thus any use of “we" in this chapter refers to Evanthia Dimara, Anastasia Bezerianos and Pierre Dragicevic.

5.1 The Attraction Effect

The attraction effect involves a *choice task* with three alternatives (i.e., Alice, Bob or Eve in our example). Alternatives are characterized by usually two attributes each (e.g. their support for education and crime control), which take values that are unambiguously ordered in terms of preference (e.g. more crime control or education is better than less). And we assume that everyone agrees that more support for education is better than less support, and similarly more support for crime control is better.

Again, an alternative *dominates* another if it is strictly superior in one attribute and superior or equal in all others (see dominance definition in chapter 2.2.1). Equivalently, Eve is dominated by Alice. An alternative is *dominated* within a set of alternatives if there is at least one alternative that dominates it. In our example there is only one dominance relation: Eve is dominated by Alice, because she is equal in education and worst in crime control. Eve is a dominated alternative, so in this choice task Eve would be formally a “wrong” answer. An alternative now is **asymmetrically dominated** within a set of alternatives if it is dominated by at least one alternative, but is not dominated by at least one other [240]. Eve is asymmetrically dominated because she is dominated by Alice but not Bob, since Eve offers better crime control than Bob. We call two alternatives **formally uncomparable** if neither dominates the other, as is the case for Alice and Bob. The best candidate is a matter of personal choice.

In a typical attraction effect experiment from the three alternatives, two that are formally uncomparable, and one that is asymmetrically dominated. They are referred to as follows: the **decoy**, the asymmetrically dominated alternative (Eve); the **target**, the alternative that dominates the decoy (Alice); the **competitor**, the alternative that does not dominate the decoy (Bob). This choice task is typically compared with a task where the decoy is absent, i.e. that involves only the two formally uncomparable alternatives.

The **attraction effect** (also called the **asymmetric dominance effect** or the **decoy effect**) is a cognitive bias where the addition of a decoy (Eve) in a set of two formally uncomparable alternatives increases people’s preference for the target (Alice) [239, 240]. In experimental settings this preference switch is observed not only for any single individual but between groups, where a higher percentage of people generally choose the target when the decoy is present. This switch in preference is irrational because it violates a basic axiom of rational choice theory, the principle of *regularity*, according to which the preference for an alternative cannot be increased by adding a new alternative to the choice set [240].

Attraction effect experiments assume that decision makers behave rationally in all other respects, and that they are able to perceive dominance relations. As a consequence, they are expected to never choose the decoy. The fact that the decoy alternative, which we never choose (since is dominated by the target), alters our choice is what makes our decision irrational.

Later on, we will generalize the attraction effect to more than three alternatives. For now, we discuss previous work on the attraction effect, which always involves two alternatives plus a decoy.

5.1.1 Why does the attraction effect occur?

Two types of explanatory theories have been offered for the attraction effect: strategic ones and perceptual ones [316].

Strategic Explanations: According to strategic theories, people use the dominance over the decoy as a heuristic to simplify an otherwise difficult decision. The face difficulty to consider attribute trade-offs between target and competitor; and to simplify it, they use information about the dominance of the target over the decoy. Choosing the target is also easier to justify to others [433] — in our example, someone who chooses Alice could argue that she is at least better than Eve. Neuroimaging studies have additionally shown that the presence of a decoy tends to reduce negative emotions associated with the choice task [218].

Perceptual Explanations: So-called "perceptual" theories assume that the addition of a decoy changes how people perceive the relative importance of the attributes involved, giving more weight to the attribute on which the target is strong [29, 219]. By analogy with perceptual contrast effects (e.g., an object appears larger when surrounded by small objects), the target appears more attractive when surrounded by unattractive alternatives [376, 435, 501]. In our example, if Eve is present, crime control may appear more important as two candidates perform relatively well on this criterion. Since this is the strength of Alice, it may raise her perceived value compared to Bob.

All explanations agree that for the attraction effect to occur, a perceptible dominance relation between the target and the decoy is key.

5.1.2 Can the attraction effect occur with visualizations?

Studies suggest the attraction effect is quite general and robust, e.g., it occurs when people choose consumer products like beers, cars, or films [240], when they gamble [500], [229], select candidates to hire [230], choose a meal in a menu [209], decide which suspect committed a crime [468], choose a policy [228] or vote [219, 358]. Even animals like hummingbirds [47], bees [424], and even brainless amoebae [295] appear to be subject to the same bias when selecting their food.

The attraction effect has been observed under a variety of experimental conditions, the majority of which present choice tasks as numerical tables including different positions for the decoy [240, 500], and the presence of unavailable alternatives [230]. At the same time, certain manipulations appear to amplify or diminish the effect. For example, the effect is smaller when there is strong prior preference or domain knowedge [336, 392], or when options are undesirable [316], but few have looked at different presentation formats.

A few studies have shown that the effect generalizes to non-tabular representations (see figures 5.2 and 5.3), such as pictures of consumer products [435], verbal instructions [422], and physical objects (i.e., people choosing between cash and a pen, or between tissues and towels) [435]. Studies have



Figure 5.2: Stimulus of attraction effect experiment with hotel photos by Frederick et al. [174]



Figure 5.3: Stimulus of attraction effect experiment with television photos by Frederick et al. [174]

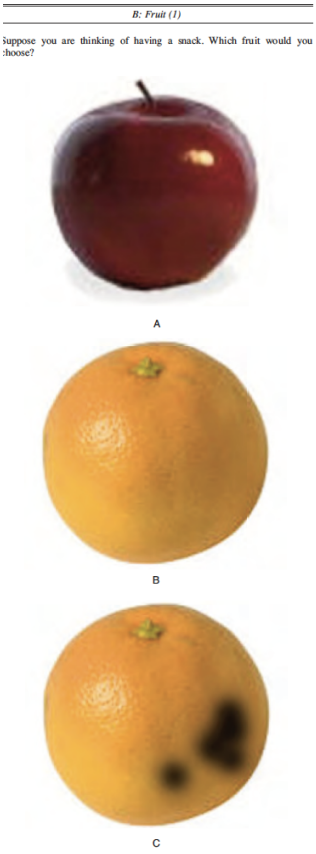


Figure 5.4: Stimulus of attraction effect experiment with fruit photos by Frederick et al. [174]

further suggested that the effect occurs when carrying out visual judgment tasks, such as finding the largest rectangle [467] or finding similarities in circle and line pairs [101].

Nevertheless, several authors [174, 523] have recently argued that the attraction effect only occurs when attributes are presented in numerical format, and reported failures to replicate the previous studies involving the representations mentioned above. Others subsequently questioned the validity of these replications [241, 434].

This debate on whether the effect generalizes to non-numerical presentations opposes *(i) numeric* displays of *quantitative* information with *(ii)* displays of *qualitative* information such as photos, verbal descriptions, or physical objects. As most data visualizations are *pictorial* displays of *quantitative* information, the debate does not provide evidence on whether the effect occurs in visualizations.

Frederick et al. [174] however studied a gambling task with two or three bets presented either as a table, or as the diagram shown in Figure 5.5 . Each bet had a prize in dollars and a probability to win. In the diagram condition, the probability of each ticket was shown as a "probability wheel" (analogous to a pie chart), and the prize was shown underneath, as a number; two gambles, a 73%

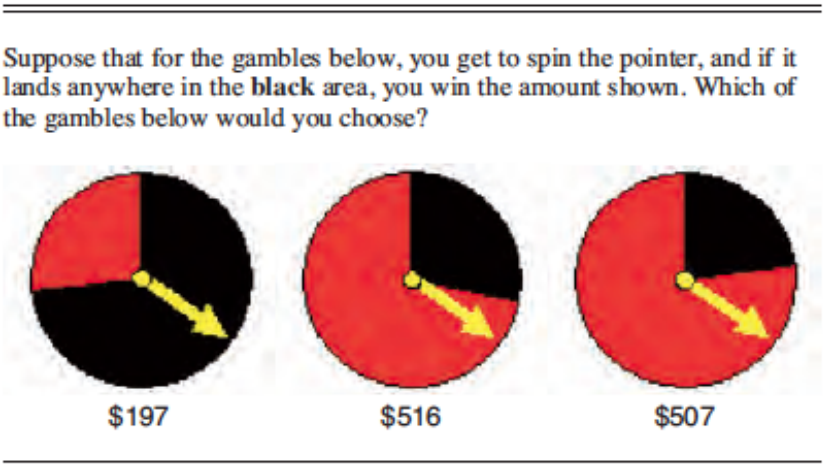


Figure 5.5: Stimulus of attraction effect experiment with gambling tickets by [174]: the probability of the ticket is shown as a "probability wheel" and the prize as a number

chance to win 12$ and 28% chance to win 33$, or three, the previous plus the decoy, a 28% chance to win 30$. When gambles were presented as numeric tables, the decoy nearly doubled the share of the target, but when pie charts were used, the effect disappeared (34% chose the target vs. 35%). To the best of our knowledge, this is the study that comes closest to a test of the attraction effect on visualizations. Nevertheless the diagram design was very domain-specific, and only one of the two attributes (probability, but not price) was encoded visually. The use of pie charts is also uncommon in professional information visualization applications [441]. We address this by using 2D scatterplots.

Although why the attraction effect occurs is still not fully understood, the possibility that it persists in visualizations is consistent with both the strategic and the perceptual explanatory theories. These theories assume that the effect requires the ability to make attribute-to-attribute comparisons and to recognize the dominance relation between target and decoy. If anything, visualizations could make these tasks easier and could perhaps even amplify the effect.

5.2 Gym Experiment: Table/Scatterplot, 3 Choices

The purpose of this first experiment is to replicate the design of a standard attraction effect experiment (two alternatives plus a decoy presented in a numerical table), and then to test if the effect persists when alternatives are shown using a scatterplot visualization.

Similar to Frederick et al. [174] who successfully replicated the attraction effect with tables but not with non-numerical formats, our study was conducted using crowdsourcing. Crowdsourcing experiments are now commonly used in information visualization [221], including in studies involving judgment and decision making [276, 332]. We used Crowdflower[1] as the crowdsourcing platform.

[1] http://www.crowdflower.com/

5.2.1 Design rationale

Although the attraction effect is thought to be robust, a replication can fail if not enough attention is paid to the details of the experimental design [241, 434]. We therefore based our design choices on lessons and recommendations from the attraction effect literature. We explain and motivate these choices here.

5.2.1.1 Factors thought to contribute to the effect

According to the literature, to maximize the attraction effect:

- The decoy should clearly appear *asymmetrically dominated*, i.e., dominated by the target but not by the competitor [241].
- The decoy should ideally be *similar to the target* [336], but different enough so that it is unlikely to be chosen [241]. For example Frederick et al. [174] the 13% of the participants chose the decoy. One possible explanation is that in the hotel dataset they used with attributes: decoration vs. price, the perception of decoration from the photo can be subjective.
- No alternative should be *highly attractive* or *highly unattractive* [217]. For example, if a product is extremely expensive and participants are not particularly wealthy, the cheap product will be chosen whether or not there is a decoy.
- Option-set: The attraction is not robust if the option-set consists of high quality products as alternatives [217],
- For similar reasons, choice tasks where participants are likely to have *strong prior preferences* about the subject matter should be avoided [336, 392]. An example would be to have flight phobia when choosing between airplanes and cars.
- The choice task should preferably be expressed as a *gain* rather than a loss [316]. For example, having to choose among below-average products can be experienced as a loss.
- The prospect of having to *justify* one's choice encourages task attention and provides a more robust effect [433].
- It is preferable to avoid *time pressure* [375].
- When the choice is *optional*, i.e. participants have the option to choose nothing, the attraction effect is stronger [131].

5.2.1.2 Dataset and scenario

By *dataset* we refer to the set of alternatives and their attribute values that make up a choice task. For example, Figure 5.1 uses the dataset $\{(5,2),(2,5),(2,4)\}$. Choosing the right dataset is critical for an attraction effect experiment since there are many cases where selecting an inappropriate dataset can reduce or eliminate the effect [434].

By *scenario* we refer to the semantic and narrative context of the choice task. In our introduction example, alternatives are candidates, attributes are support for education and crime control, and the

decision consists of voting for a candidate.

Many different scenarios and attribute values have been employed since the original studies of the attraction effect [239, 240]. We reasoned that a recent study is more likely to employ an optimal design, since it has more accumulated knowledge to build on. We therefore chose to replicate the scenario from the first experiment of recent work by Malkoc et al. [316], that involved choosing a fitness club (or gym), and found a clear attraction effect. The experiment's design fulfills most of the criteria from the previous subsection, and a clear effect was found in the original study.

In Malkoc et al.'s study, each gym was defined by its variety and its cleanliness, both rated from -10 to +10. A positive rating meant better than average, and a negative rating meant worse. The study investigated whether undesirable options (all negative ratings) eliminate the attraction effect. But as the effect was strong for their control condition (all positive ratings), we chose it for our replication.

The experiment employed four gyms g_C, g_V, g_C^*, g_V^*, where g_C was cleaner, g_V had more variety, and g_C^* and g_V^* were slightly less attractive than g_C and g_V respectively. The attribute values were $g_C(\textit{variety}\ 1, \textit{cleanliness}\ 4)$, $g_V(4,1)$, $g_C^*(0,4)$, and $g_V^*(4,0)$. Three choice tasks were tested: $\{g_C, g_V\}$ (no decoy), $\{g_C, g_V, g_C^*\}$ (decoy on g_C), and $\{g_C, g_V, g_V^*\}$ (decoy on g_V). These attribute values however cause the data points g_C^* and g_V^* to overlap with scatterplot axes, possibly creating visual anchoring effects that could affect participant responses. Since such effects were outside the scope of our study, we incremented all values by one. Thus we used as attribute values $g_C(2,5)$, $g_V(5,2)$, $g_C^*(1,5)$, and $g_V^*(5,1)$. These values preserve all dominance and similarity relationships between alternatives, the target, the competitor and the decoy.

5.2.1.3 Stimuli: tables and scatterplots

In order to test the attraction effect in visualizations, we need to first confirm that the attraction effect exists indeed in a tabular form for our specific dataset, due to the possible factors mentioned above that can reduce or eliminate the effect. We also used a numerical table as a control condition, to test our experiment design and compare our results with previous studies. Figure 5.6a shows the 2×2 table representation for the multi-attribute choice task $\{g_C, g_V\}$, and Figure 5.6b shows the 3×2 table for the choice task $\{g_C, g_V, g_V^*\}$. Attributes were presented in rows and alternatives in columns, as in Malkoc et al. [316]. Alternatives were labeled A, B or A, B, C from left to right. The ordering of rows and columns in the table will be discussed in the next subsection.

In the visualization condition, alternatives were conveyed with scatterplots (see Figure 5.6c,d) and similarly labeled A, B or A, B, C from left to right and from top to bottom.

There are four main reasons behind the choice of scatterplots for the visualization condition. First, 2D scatterplots are a standard information visualization technique [156, 347]. Second, they are suited for visualizing any tabular dataset with two quantitative dimensions, which captures the choice tasks used here and most choice tasks used in previous studies on the attraction effect. Third, a scatterplot shows all data cases within the same frame of reference, thus providing a rapid overview of all alternatives. A unified frame of reference also likely supports comparisons better than side-by-side

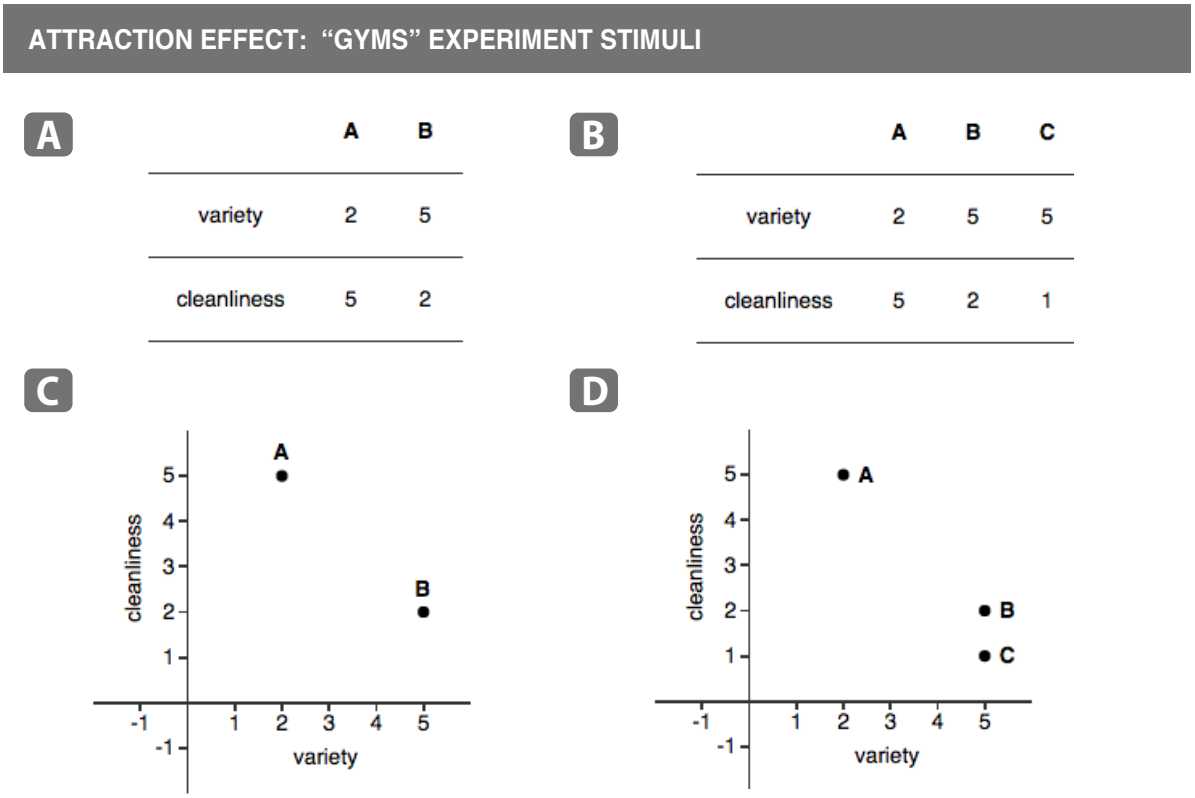


Figure 5.6: Examples of experimental stimuli for the table (a,b) and the scatterplot (c,d) conditions. The left multi-attribute choice task (a,c) has no decoy, while the right choice task (b,d) has a decoy on B. A and B are the incomparable alternatives, and the decoy C is in this example on the target B. Presentation in the experiment: size 250x260 cm, font Helvetica, 15px

views such as Frederick et al.'s [174] pie charts discussed in the background section. In fact, scatterplots are used as figures in most articles on the attraction effect for conveying the alternatives used in the experiments [47, 131, 217, 228–230, 239, 240, 295, 316, 336, 358, 375, 376, 424, 433]. Finally, scatterplots scale up to more than three items, which is an important requirement for our follow-up experiment.

The appearance of tables and scatterplots was kept as similar as possible to avoid experimental confounds due to choices in visual design. Both presentation formats took approximately the same amount of screen real estate, and graphical attributes (colors, line thickness and font sizes) were kept consistent. In both conditions, participants indicated their choice through separate radio buttons.

5.2.1.4 Ordering of alternatives and attributes

Although Malkoc et al. [316] used a fixed order of presentation for attributes and alternatives, the choice of ordering may affect participant responses. This is not a problem with regular attraction effect experiments, but could potentially bias our results because the two formats vary in the ways alternatives and attributes can be ordered, and in the ways ordering could affect choices. For example,

participants may give more weight to variety if it is shown first on a table, but on a scatterplot, it is not clear whether the choice of horizontal vs. vertical axis would have a similar effect. In addition, alternatives can be presented in any order within a table, while on a scatterplot the way alternatives are laid out is dictated by attribute values.

To balance out any possible order effect, we thus randomized the order of presentation of attributes and alternatives across participants. In the scatterplot condition, axes can be flipped, leading to 2 different attribute orderings (variety on x and cleanliness on y, or vice versa). In a 2×2 table, there are 2 ways to order rows and 2 ways to order columns, yielding 4 different tables. Similarly, a 2×3 table can be presented in 12 different ways. Since the decoy is typically placed next to the target in attraction effect experiments (e.g., [174, 217–219]), we removed cases where the target was not next to the decoy (4 tables out of 12). Since the decoy cannot appear between the target and the competitor in the scatterplot, we also removed cases where the decoy was in the middle (4 tables out of 12). In summary, we used 4 different table stimuli and 2 different scatterplot stimuli for each of the three choice tasks $\{g_C, g_V\}$, $\{g_C, g_V, g_C^*\}$ and $\{g_C, g_V, g_V^*\}$, for a total of 18 different experimental stimuli.

5.2.1.5 Crowdsourcing quality control

When designing a crowdsourcing experiment we need to make sure that the participants will perform the task, will pay enough attention and understand the task. Quality control is important in any crowdsourced experiment [221], and in attraction effect studies in particular [434]. Quality was ensured by recruiting highly-rated crowdsourced contributors (level 3 on the Crowdflower platform), by including test questions, and by devising a job assessment scheme prior to running the experiment. Four criteria were used for job assessment:

Duplicates. As the Crowdflower platform had no direct way to prevent a participant from repeating a job, we explicitly stated that we would not accept duplicate participations, and identified duplicates by contributor IDs.

Completion time. A job completion time of less than 1 minute or more than 30 minutes was considered abnormal. Our pilots indicated an average task completion time of 6 minutes.

Justification. Participants had to provide a textual justification for their choice. Justifications were classified by one investigator as either proper or improper, depending on whether it made a reference – direct or indirect – to either cleanliness or variety. We informed participants in advance that they would have to justify their answer, both for fairness and for better results (see Section 5.2.1.1).

Prior preferences. After the experimental task, participants were asked if they suffered from an abnormal fear of dirt (or bacteriophobia), with "no", "yes", or "unsure" as answers. This identified participants with a strong prior preference for cleanliness, as strong prior preferences are known to reduce the effect [336, 392] and thus were likely to be insensitive to the attraction effect in this particular case (see Section 5.2.1.1).

Table and scatterplot tests. After carrying out the task, participants were subjected to two screening tests: a *numerical table test*, and a *scatterplot test*, irrespective of the condition they saw. Both

tests involved choosing between three laptops based on their RAM and CPU, with one laptop clearly dominating the other two (i.e., had both higher RAM and higher CPU). The tests were designed to be trivial, with a single correct answer, using a presentation format similar to the experimental task (see Figure 5.6). The purpose of the table test was to screen for contributors who did not pay attention to the tasks. The purpose of the scatterplot test was to control for visualization literacy [65], and make sure that participants were able to read scatterplots and to perceive dominance relations [241].

Our job assessment scheme classified jobs in three categories: the *Red*, where the job is rejected (and the contributor not paid); the *Orange*, where the job is accepted but the data discarded from our analysis; and the *Green*, where the job is accepted and the data kept in our analysis. Due to limitations in the Crowdflower platform we had to pay all contributors, but we report here on the three categories nonetheless.

A total of 437 jobs were submitted, after removing invalid completion codes and duplicate worker IDs. A job was marked *Red* if: the completion time was abnormal (1 % of all submitted jobs), the gym choice was not properly justified (14%), or the contributor failed the table test (12%). A job was marked *Orange* if: the response to the bacteriophobia question was "yes" (12% of all submitted jobs), or the contributor failed the scatterplot test (13%). In total, 16% of all submitted jobs were marked *Red* and 14% were marked *Orange*. These jobs were discarded from all our analyses.

5.2.2 Experiment design

The experiment followed a 3×2 between-subjects design. The first independent variable was the multi-attribute choice task, which involved three different datasets: $\{g_C, g_V\}$, referred to as the *no decoy* condition; $\{g_C, g_V, g_C^*\}$, referred to as *decoy on cleanliness*; and $\{g_C, g_V, g_V^*\}$, referred to as *decoy on variety*. The second independent variable was the presentation format, with two conditions: *table* and *scatterplot*.

5.2.2.1 Procedure

We conducted a first pilot study to ensure the clarity of the instructions, and we then uploaded the experiment as a Crowdflower job.

Participants had to open an external 8-page Web form. They were told they would have to choose a fitness club based on two attributes: variety of the machines and cleanliness of the club. They had to assume that they had done some preliminary research, and had narrowed down their choices to two (in the no-decoy condition) or three (in the decoy conditions) clubs. They were then shown the gyms as a table or a scatterplot (Figure 5.6) and asked to choose one.

Once finished, participants rated their confidence on a 7-point scale and provided an open text justification for their choice. They also rated their enthusiasm towards fitness clubs on a 7-point scale and reported on whether they suffered from bacteriophobia, an abnormal fear of dirt, as mentioned previously. Finally, they were given the table and scatterplot tests (Section 5.2.1.5), and filled a short questionnaire with demographic information. The navigation mechanism forced the completion of

each question before proceeding to the next page. Participants could review previous pages, but not change their answers.

At the end of the experiment, participants copied the provided completion code and pasted it in the crowdflower platform to receive payment. The entire job took on average 6 minutes to complete, and participants were paid $0.60 upon completion.

5.2.2.2 Participants

Our population sample consisted of 305 crowdsourced contributors who submitted valid responses, i.e., jobs classified as *Green* (Section 5.2.1.5). Job assignments were left on the crowdsourcing server until the planned sample size of n=50 per condition was approximately reached. [2] So as not to violate the principle of random assignment, we decided against targeting n=50 precisely. We obtained n=54, 51, 50 for the *table* choice tasks, and n=47, 53, 50 for the *scatterplot* tasks.

A summary of our participants' self-reported demographics is shown in Figure 5.7 (map and bar charts labeled "Gyms"). As can be seen, participants tended to be educated young male adults.

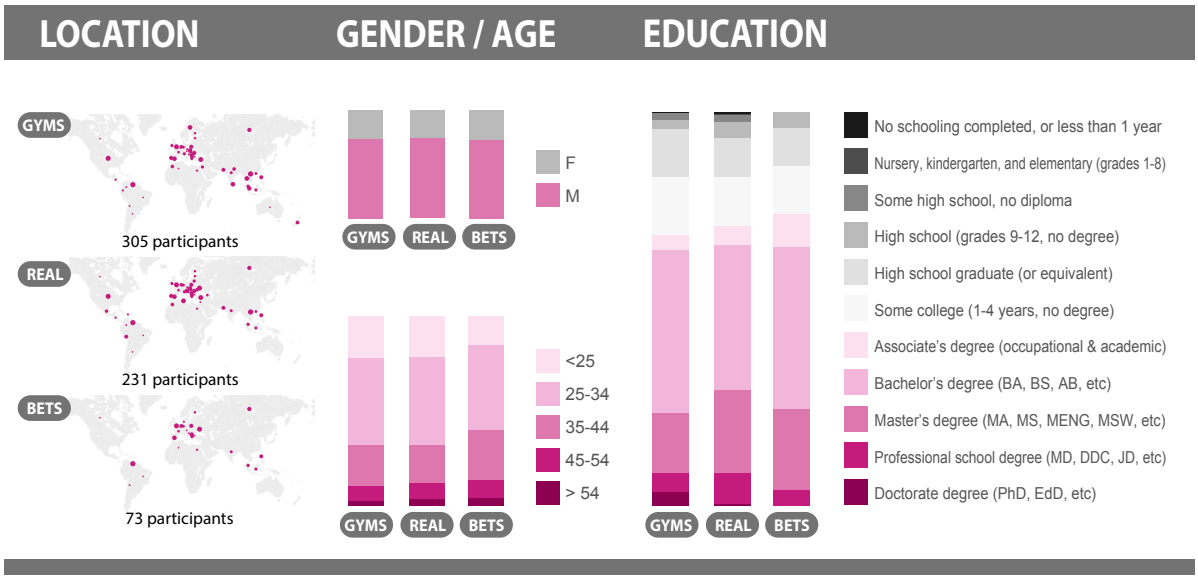


Figure 5.7: Participant demographics for both experiments.

5.2.2.3 Hypotheses

Following our discussion in the section 5.1 our research hypotheses were:

Hr1 When a choice task is presented as a numerical table, the addition of a decoy increases the attractiveness of the target.

[2] Some conditions took longer to reach n=50 due to random fluctuations in the occurrence of invalid jobs, leading to a situation where only a subset of our conditions remained available on the crowdsourcing system.

Hr2 When a choice task is presented as a scatterplot, the addition of a decoy increases the attractiveness of the target.

These translate into the following statistical hypotheses:

H1 A larger proportion of participants will choose the target in the *table* × *decoy on cleanliness* and the *table* × *decoy on variety* conditions than in the *table* × *no decoy* condition.

H2 A larger proportion will choose the target in the *scatterplot* × *decoy on cleanliness* and the *scatterplot* × *decoy on variety* conditions than in the *scatterplot* × *no decoy* condition.

5.2.3 Results

As before, we analyze, report and interpret all our inferential statistics using interval estimation [140]. The experimental stimuli, data and analysis scripts are available at `http://www.aviz.fr/decoy`.

5.2.3.1 Planned analyses

All analyses reported in this section were planned before data was collected. One planned analysis (an analysis of differences between attraction effects) was not conducted because it required equal sample sizes across all conditions.

Only one participant out of 306 chose a decoy, which is low compared to previous studies, where decoy selection rates can be as high as 13% [174]. This shows that participants carried out the tasks seriously and could perceive dominance relationships. The decoy choice is removed from the rest of this analysis.

Participant choices are shown in the top of Figure 5.8 marked "Gyms" ("Real" and "Bets" refers to our second and third experiment respectively) .The top three bars are for the *table* format, in the conditions *no decoy*, *decoy on cleanliness* and *decoy on variety*. Adding a decoy is expected to increase the proportion of choices of the target, in the direction indicated by the arrow. This was indeed the case for the *decoy on variety* condition (a 20% increase), but not for *decoy on cleanliness* (a 6% decrease). The next three bars refer to the *scatterplot* format. Here the expected increase was observed for both *decoy on cleanliness* (a 18% increase) and *decoy on variety* (a 3% increase). We now turn to inferential statistics to determine to what extent these effects are reliable.

The previously reported effects are shown in Figure 5.9 — the four black dots under the category “Gyms”. Effects are expressed in percentage points, where a positive value (i.e., to the right of the vertical dashed line) indicates an attraction effect. Dots are sample statistics, while error bars are 95% confidence intervals indicating the range of plausible population effects [123]. Confidence intervals were computed using score intervals for difference of proportions and independent samples.

Figure 5.9 shows that the unexpected reversal observed in the *table* × *decoy on cleanliness* is too unreliable for any conclusion to be drawn. The same is true for the small effect found for *scatterplot* × *decoy on variety*. However, we have good evidence for an attraction effect in the other two conditions. The magnitude of the effect is comparable to Malkoc et al. [316], shown on the top of Figure 5.9.

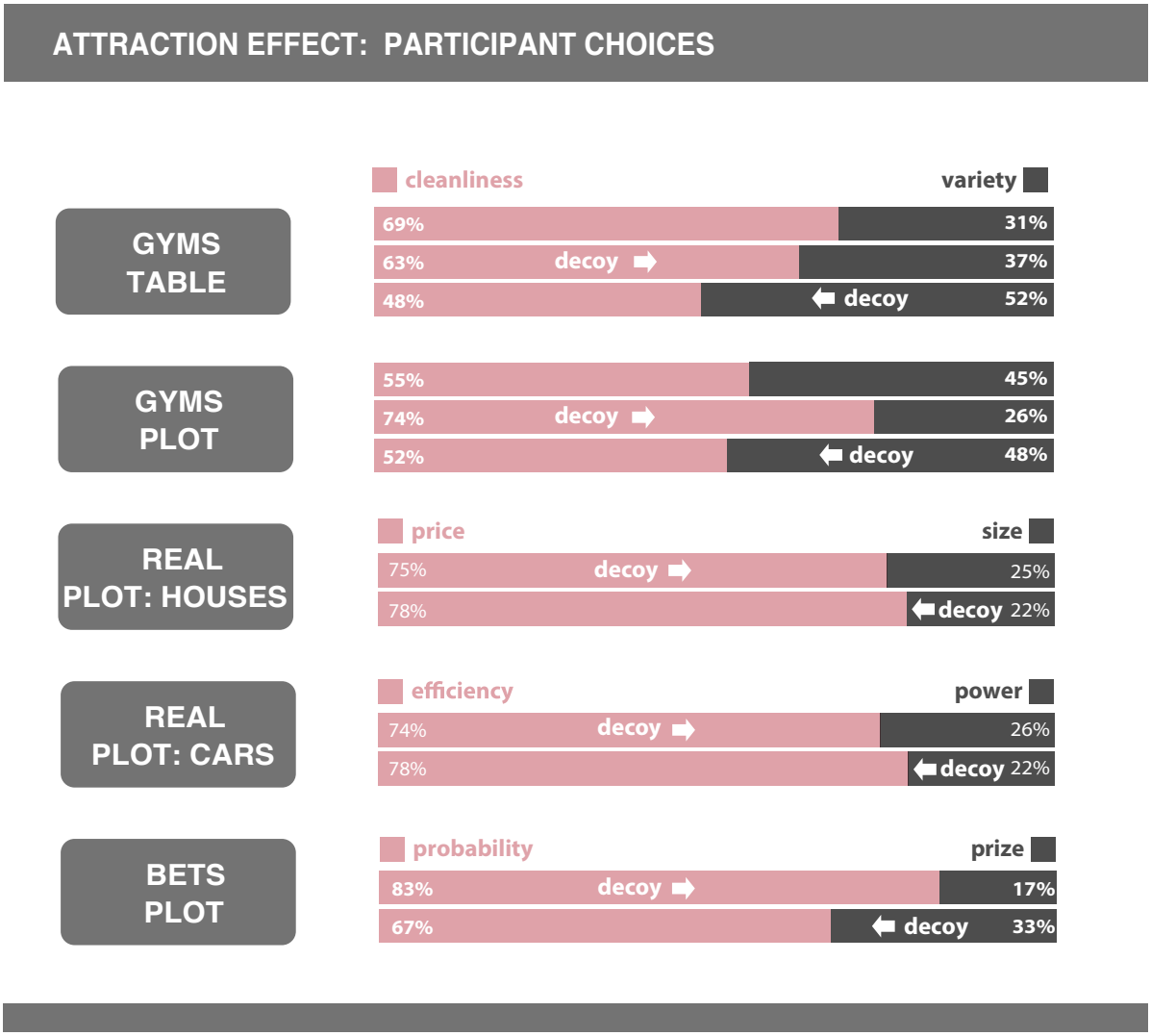


Figure 5.8: Proportions of participant choices in all three experiments.

Thus, our results partially confirm **H1** and **H2**, but are less "clean" than in Malkoc et al.'s [316] original study.

5.2.3.2 Additional analyses

Participants reported similar confidence in their answers across all conditions (Figure 5.10). They were overall highly confident, with a mean rating of 5.9 to 6.1 on a 7-point Likert scale, depending on the condition. Participants' reported familiarity with fitness clubs varied, but they were overall rather familiar (Figure 5.10).

We computed *combined attraction effects*, shown as pink dots and error bars in Figure 5.9. A combined attraction effect is the sum of the attraction effects obtained in both decoy conditions, or equivalently, the difference in choice proportions between these two conditions (i.e., the differences



Figure 5.9: Point estimates and 95% confidence intervals for the attraction effects in Malkoc et al. [316], and in our two experiments.

between the two left bars or the two right bars across two "decoy" conditions in Figure 5.8). This combined measure generally yields more statistical power and facilitates comparisons of results since some experiments (e.g., [500] and our next experiment) do not include a no-decoy condition and thus only report combined attraction effects.

The two pink error bars in Figure 5.9-Gyms show that the data overall speaks in favour of an attraction effect, both for the table and the scatterplot. To better quantify the strength of evidence, we conducted a Bayesian analysis using the Jeffreys prior for proportions [60]. Ignoring previous studies



Figure 5.10: Self-reported confidence and familiarity in all three experiments ("Gyms", "Real" and "Bets").

and considering our data only, the presence of a combined attraction effect in the *table* condition is 34 times more likely than a practically null effect (set to $\pm 1\%$), and 11 times more likely than a "repulsion" effect. In the *scatterplot*, a combined attraction effect is 150 times more likely than a practically null effect, and 66 times more likely than a repulsion effect.

5.2.3.3 Discussion

We found evidence for an attraction effect on *table* for the *decoy on variety* condition, but not for the *decoy on cleanliness* condition, where the effect may be smaller or even possibly negative (see Figure 5.9). We do not have an explanation for this asymmetry, but the wide confidence intervals and their large overlap suggests that the difference may be due to a large extent to statistical noise [123] .

Based on the combined attraction effect which is a more holistic measure with more statistical power, we replicated the attraction effect on tables (**H1**) but the results are less strong than in the initial study [316] (i.e., about half of the original study, as shown by the pink CIs in Figure 5.9). It is common for a replication to yield smaller effect sizes [108], but the differences in results could also be due to modifications we made to the original experiment design.

We produced four different stimuli for each choice task in order to eliminate possible presentation order effects for alternatives and attributes, whereas Malkoc et al. [316] used a unique table. The use of different stimuli could have yielded a higher variability in responses.

Our study was also a crowdsourced experiment, whereas Malkoc et al. conducted theirs with students in a lab, where participants are less diverse and generally more focused [332]. Perhaps the feeling of being evaluated was also stronger for students, which we know can amplify the attraction effect [433]. Our rejection criteria (e.g., textual justification for the answer, table and scatterplot test, attention test) could have also filtered subsets of the population that are more vulnerable to the effect. Finally, our participants were on average rather familiar with gyms (Figure 5.10), and 11% were unsure if they suffered from bacteriophobia, and we know that familiarity with the subject matter and strong prior preferences can reduce the effect [336, 392]. Malkoc et al. [316] do not report on familiarity and prior preferences.

Despite mixed results for the table condition, we obtained good evidence for an attraction effect in the *scatterplot* condition. There still appears to be an asymmetry between the two decoy conditions (this time, in the opposite direction), but CIs show no evidence for a difference. The combined attraction effect provides compelling evidence that the attraction effect can generalize to scatterplots (**H2**). This observed shift in preference after adding an irrelevant option to a two-point scatterplot gives credence to the idea that people may make irrational decisions even when they use visualizations as decision making aids. Thus we decided to explore the effect further, using scatterplots with larger sets of alternatives.

5.3 Extending the Attraction Effect

Our gym experiment confirmed that the attraction effect can extend to scatterplot formats. However, we have so far only considered three data points, which does not capture most real-word decision tasks where visualizations would be used.

5.3.1 Can the attraction effect generalize to many alternatives?

Previous work has focused on only three alternatives because in numeric tables, it is hard to perform rapid attribute-to-attribute comparisons and recognize dominance relationships between many points. Bettman et al. [59] point out that the attraction effect requires asymmetric dominance relationships to be "perceptual in nature" and "easy to access". They expect that the bias will be eliminated with multiple alternatives, as the number of pairwise comparisons increases and these relationships become

harder to understand. This may be true for numerical tables, but not necessarily for visualizations such as scatterplots, that are designed to aid viewers read and understand complex data, and support comparison of many data points at once [347].

It is thus plausible that visualizations of many alternatives can also elicit attraction effects.

Thus, we expect that in scatterplots, *multiple* dominated datapoints (high number of decoys) will also increase the attractiveness of their dominant datapoint. We decided to test this extension of the attraction effect in a multiple datapoint scatterplot.

5.3.2 Ways of adding more alternatives

There are three ways the classical attraction effect procedure can be extended to include more than three alternatives:

1. By adding more non-dominated options. In our introduction example, the only non-dominated alternatives were Bob and Alice. We could add more candidates that neither dominate nor are dominated by Bob and Alice. The set of formally uncomparable or non-dominated alternatives is also called the *Pareto front*.
2. By adding more decoys. In our example the only decoy is Eve. We could however add more decoys similar to Eve.
3. By adding "distractors", i.e., irrelevant options that play neither the role of target, of competitor, or of decoy. An example would be a dominated candidate that appears both in the baseline condition and in the decoy condition.

The first approach is problematic in at least two respects. One is that since it breaks the dichotomy between target and competitor, it would require a major change in the way the attraction effect is measured in experiments. A second problem is that it would cause the attraction effect to interfere with other cognitive biases. For example, the *compromise effect* is a bias by which if presented with several formally uncomparable alternatives, people tend to avoid extremes and choose options in the middle [433]. Even though it could be informative to study how the two effects may combine, we decided here to focus on the attraction effect only.

Adding an arbitrary number of distractors (option 3) is however possible. With many distractors a single decoy is unlikely to produce a measurable effect, but more decoys can be added (option 2). The Pareto front however still needs to consist of only two alternatives – a target and a competitor. We present an extension of the attraction effect procedure using this approach.

5.3.3 Extended procedure

The procedure consists of starting with a **baseline choice task** T_0 (see middle of Figure 5.11 for an example). This baseline choice task has two non-dominated alternatives, A and B. All other alternatives are dominated by A and/or B, and are called **distractors**.

For convenience, we divide the space of all possible alternatives into three dominance regions, shown in Figure 5.11. If d_A is the region dominated by A (blue hatches in the Figure) and d_B is

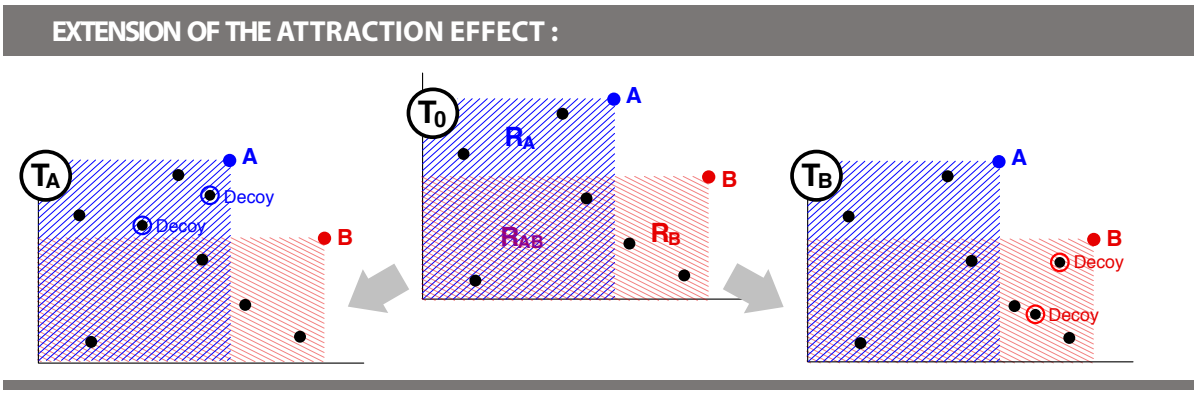


Figure 5.11: A baseline choice task T_0 and two possible test choice tasks: T_A, where A is the target, and T_B, where B is the target.

the region dominated by B (red hatches), then R_A $d_A \setminus d_B$ (region dominated by A but not by B), R_B $d_B \setminus d_A$ (region dominated by B but not by A), and R_{AB} $d_A \cap d_B$ (region dominated by both). In the figure, the baseline choice task contains two distractors per region.

From the baseline choice task one can derive two types of test choice tasks, labeled T_A and T_B in the figure. The choice task T_A is created by adding extra alternatives to the region R_A. Thus T_A only differs from T_0 in that it contains more alternatives that are dominated by A but not by B (twice as many, in this example). These *extra* asymmetrically dominated alternatives are referred to as **decoys**, while A is called the **target** and B the **competitor**. Similarly, the task T_B is created by adding extra decoys to the region R_B, and this time B is the target and A is the competitor.

In case no distractor is included in T_0 and a single decoy is added to T_A and to T_B, we obtain a classical attraction effect experiment. Thus our new definitions for decoy, target and competitor are consistent with the definitions from Section 5.1 and generalize them to more complex cases. However, decoys, targets and competitors are always defined with respect to a baseline choice task.

5.4 Real Experiment: Scatterplot, Many Choices, Realistic dataset

Real world datasets have hundreds and even thousands of data points, and they, or parts of them, are often rendered in visualization software. Visualizations such as scatterplots, are designed to aid viewers read and understand complex data quickly. Nevertheless, it is possible that this ease of reading and comparing data can lead to attraction effects when dealing with multiple datapoints.

5.4.1 Design rationale

In the first experiment **"Gyms"**, we found evidence for an attraction effect in scatterplots by replicating a standard experimental protocol. However, the datasets are limited to 2 or 3 alternatives, which is not realistic for a dataset people may want to visualize. The main reason for this limitation of previous work

is that in numerical table representations it is hard to perform rapid attribute-to-attribute comparisons, and thus recognition of dominance relationships, between many alternatives.

Scatterplots and other visualizations are designed to remove these barriers, and facilitate rapid spatial comparisons of many data points. Thus, we hypothesized that in scatterplots with many alternatives, a sufficient number of decoys will also increase the attractiveness of the target. We decided to test this in a second experiment, using sets of alternatives derived from two real datasets. In this section we describe the motivation, design rationale and results of this experiment, referred to as experiment **"Real"**.

5.4.2 Terminology

Here we extend the terminology defined in subsection 5.3.3.

The **Pareto front** is the set of all alternatives that are not dominated in a choice task [363]. In other words, it is the set of all possible choices that are not obviously "wrong". All alternatives in a Pareto front are formally uncomparable. In a classical attraction effect experiment (Figure 5.1), the Pareto front consists of only two alternatives: the target and the competitor.

As we explained insubsection 5.3.3 , T_1 is a choice task where only two choices A and B are available on the Pareto front. T_2 is another choice task that only differs from T_1 in that it contains additional alternatives, all dominated by A but not by B. We refer to these additional alternatives as **decoys** on A, to the alternative A as the **target**, and to the alternative B as the **competitor**. Concrete examples of such cases will be provided later on.

A **constrained choice task** is a choice task that requires choosing an option from a subset of all the alternatives. We will refer to this subset as the **choice set**. A classical attraction effect experiment is not a constrained choice task, since the choice set is the same as the set of alternatives.

5.4.3 Stimuli and task

As in Gyms experiment, we chose a 2D scatterplot to visually represent the sets of alternatives, one of the common infovis representations of large bi-dimensional datasets. This time we did not include a numerical table as a control condition, since numerical tables do not support rapid comparisons among many alternatives. Instead we looked at how the addition of decoys shifts participants' choice between target and competitor.

In numerical tables or other textual formats, people need to actively perform pairwise comparisons to see dominance relations between alternatives. According to [**?**], the choice task in an attraction effect experiment can be broken in two stages: *i)* a dominance recognition stage, where participants exclude the dominated alternative; and *ii)* a final selection stage, where participants choose one of the two non-dominated alternatives based on their preference.

Our concern regarding the extension of attraction effect in scatterplots was that the dominance recognition task – i.e. the exclusion of multiple decoys– could be time consuming or error-prone for crowdsourced participants, given the large number of alternatives. Since we were interested in

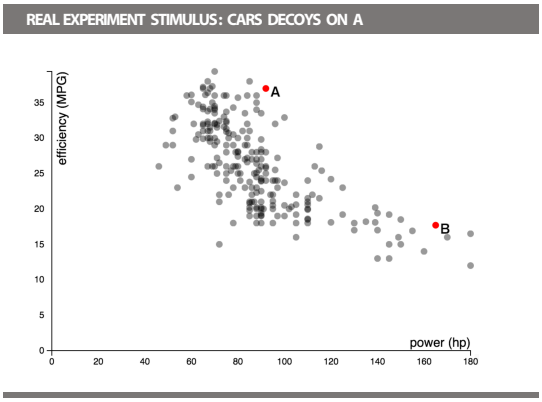


Figure 5.12: Cars dataset, decoys on A

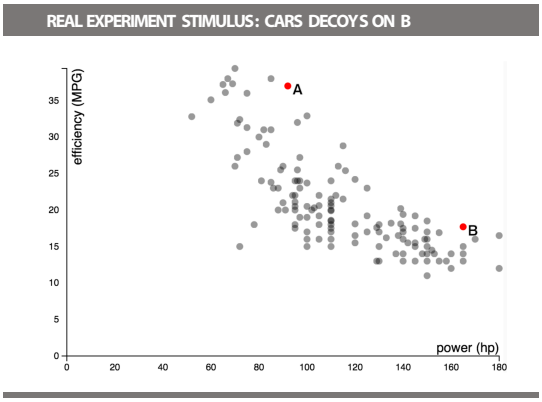


Figure 5.13: Cars dataset, decoys on B

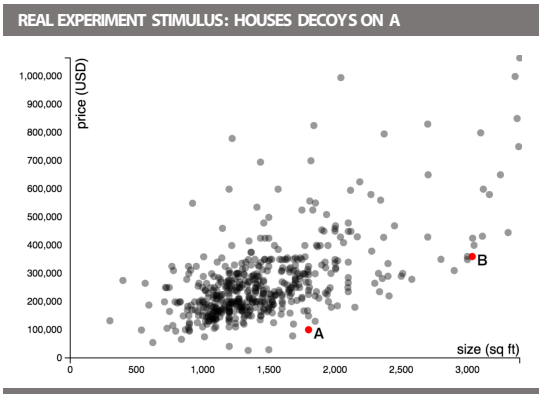


Figure 5.14: Houses dataset, decoy on A

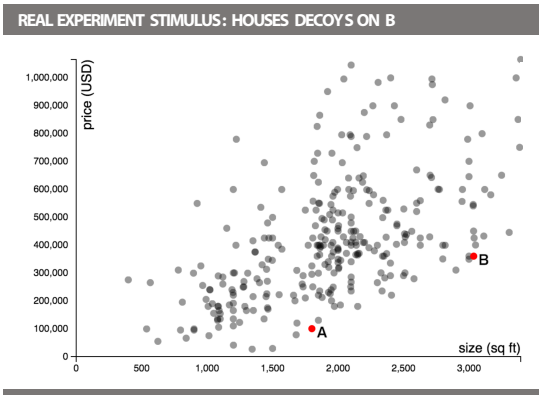


Figure 5.15: Houses dataset, decoy on B

the selection stage of the task, we decided to eliminate the dominance recognition part of the task by restricting participants' choice set to two non-dominated alternatives. We focused on only two alternatives instead of the full Pareto front to remain consistent with the target/competitor dichotomy used in studies on the attraction effect and to avoid interaction with other cognitive biases, as explained in [134].

Thus, we used a *constrained choice task* whose the *choice set* consisted in two formally uncomparable alternatives picked on the Pareto front. As we can see in the stimuli shown in Figures 5.12, 5.13, 5.14, and 5.15, we indicated target and competitor in red color –whereas other data points were in gray– and we labeled them as A and B. Participants indicated their choice of A and B in a radio button below the diagram.

5.4.4 Scenario and attribute values

Previous research has studied how different positions of a single decoy influence the effect, and it is believed that positions are not all of the same weight. For instance, [240] found the attraction effect increasing in the dimension on which the target is the weakest (i.e., the dimension where the

competitor is superior). Despite the differences in attraction intensity depending on position, the effect seems to persist regardless of where we place a single decoy [500]. However, the decoys so far have been studied in one position at a time (single decoy), rather than having multiple decoys, at different positions at the same time, as is our case.

5.4.4.1 Dataset selection

Since we are interested in the attraction effect with many alternatives shown on scatterplots, for this second experiment we based our tasks on real datasets. We started with a corpus of five multidimensional tabular datasets containing information on houses, cars, cameras, cereals and movies. These datasets are routinely used in the infovis community for the purposes of teaching, designing and demonstrating multidimensional data visualization systems (e.g., [156, 525]).

For each dataset we examined all possible 2D scatterplots and searched for distributions that were roughly linear (such as in Figure 5.16), indicating a trade-off between the two attributes. In addition, each attribute should be easy to understand, and, should exhibit a clear direction of preference. For example, *carbohydrate content* from the cereals dataset is not a good choice of attribute, since some people might seek low carbohydrate content while others might seek the opposite, and some may not even understand what it means. Only two datasets (cars and houses) had attributes pairs that had a roughly linear relationship and where each attribute was easy to understand and had clear direction of preference. We chose a pair of attributes for each of these datasets. More specifically, the two bi-dimensional datasets we extracted from our corpus were:

- The *cars* dataset, a set of 407 car models from America, Japan and Europe manufactured from 1970 to 1982, described according to their *horsepower* and their fuel *efficiency* in miles per gallon.
- The *houses* dataset, a set of 781 real estate listings in the San Luis Obispo area from 2009, described according to their *size* in square feet and their *price* in US dollars.

Although the car and real estate markets have significantly evolved since the data was collected, the trade-offs involved remain the same and thus choice tasks should not be impacted.

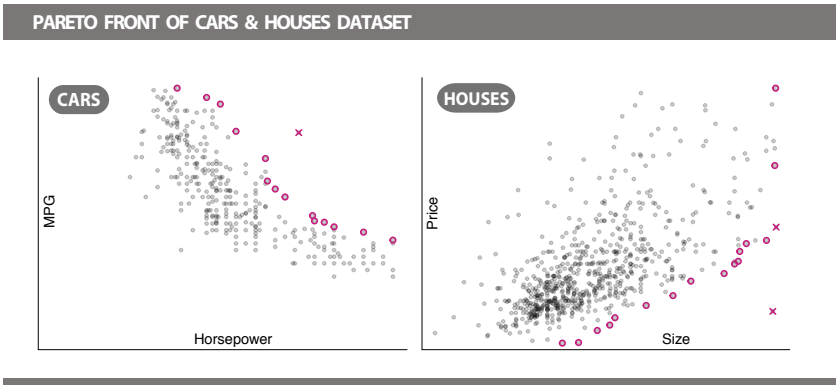


Figure 5.16: The *cars* and *houses* datasets with their Pareto front, in pink. Crosses are discarded outliers.

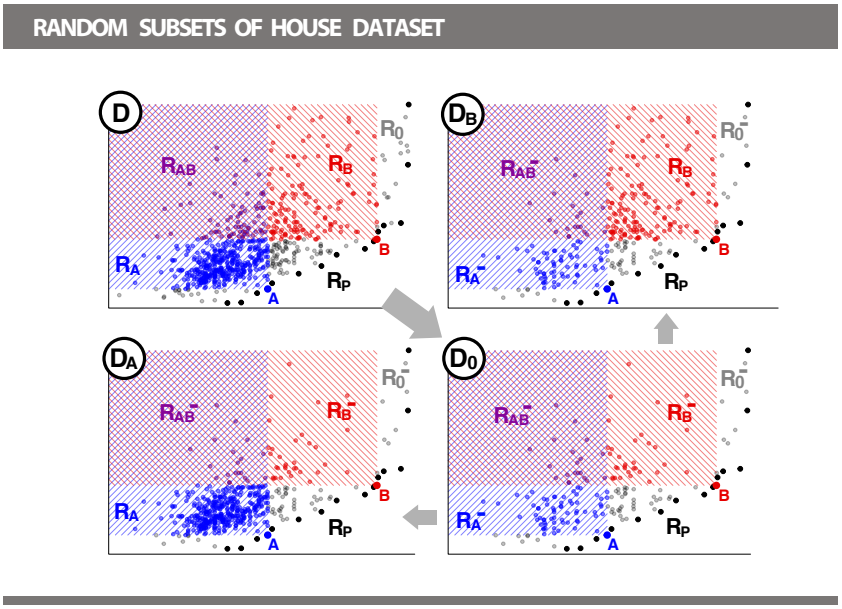


Figure 5.17: The house dataset D and three possible subsets (D_0, D_A and D_B) created by randomly removing alternatives in specific regions.

5.4.4.2 Dataset preparation

As explained in Section 5.3.2 we extended the classical attraction effect procedure by creating 'distractors", i.e., irrelevant options that play neither the role of target, of competitor, or of decoy.

We cleaned up the chosen datasets by *i)* removing all duplicate alternatives — i.e., houses or cars with identical attribute values, and *ii)* removing alternatives more than two standard deviations away from the mean on either of their two attributes. This prevented outliers from excessively compressing the scales of the scatterplot axes. A total of 61 duplicates and 37 outliers were removed from the car dataset, and 16 duplicates and 43 outliers were removed from the house dataset.

We then removed outliers along the Pareto front on each of the two datasets. This was done by performing a linear regression on the Pareto front and removing all alternatives whose standardized residual was greater than two. Doing so eliminated alternatives that may appear attractive only because they present an unusually good trade-off. One such outlier was removed from the car dataset, and two from the house dataset. Figure 5.16 shows the two datasets, their Pareto front, and the discarded outliers.

We then chose two alternatives A and B on the Pareto front to act as target and competitor. The alternatives A and B were chosen so as to maximize $|R_A|$ $|R_B|$ as defined below.

We then generated three subsets from each dataset as follows

Let D be the full set of alternatives (cars or houses), P its Pareto front, $\{A, B\}$ the choice set, d_A all alternatives dominated by A, and d_B all alternatives dominated by B (see blue and red hatchings respectively in Figure 5.17 Ⓓ). We partitioned D in five regions: R_A $d_A \setminus d_B$; R_B $d_B \setminus d_A$; R_{AB} $d_A \cap d_B$; R_P P; and R_0 $D \setminus (d_A \cup d_B \cup P)$. We then randomly eliminated 80% of the alternatives from the

sets R_A, R_B and R_{AB}, and 50% of the alternatives in R_0, yielding the subsets R_A^-, R_B^-, R_{AB}^-, and R_0^-. From this we constructed three sets of alternatives, shown in Figure 5.17: $D_0 \; R_A^- \cup R_B^- \cup R_{AB}^- \cup R_0^- \cup R_P$; $D_A \; R_A \cup R_B^- \cup R_{AB}^- \cup R_0^- \cup R_P$ and $D_B \; R_A^- \cup R_B \cup R_{AB}^- \cup R_0^- \cup R_P$.

It can be seen in Figure 5.17 that the only difference between D_0 and D_A is that D_A contains more alternatives that are dominated by A but not by B. These asymmetrically dominated alternatives are analogous to *decoys*, while A plays the role of a *target* and B plays the role of a *competitor*. Note however that no artificial choice is added: all points belong to the original dataset, including the decoys. The roles of A and B are swapped when comparing D_0 to D_B. The final stimuli can be seen in Figures 5.12 to 5.15.

5.4.5 Measures

We measure the attraction effect by the difference in the proportion of participants who chose the target in the condition with decoys vs. the condition without decoys. If $p(X)_S$ is the proportion of participants who chose X in the set of alternatives S, our data subsets allow for two possible measures of the attraction effect: $E_A \; p(A)_{D_A} - p(A)_{D_0}$ and $E_B \; p(B)_{D_B} - p(B)_{D_0}$. A third aggregated measure $E_{AB} \; E_A \; E_B$ referred to as the *combined attraction effect* is possible that combines the two effects and has been used in past studies [500]. Since $E_{AB} \; p(A)_{D_A} - p(A)_{D_B}$, the combined measure can be calculated without knowing the responses for D_0. To maximize statistical power, we therefore chose to only present the choice tasks D_A and D_B, and use E_{AB} as the measure of attraction effect. We report this measure for both the *cars* and the *houses* dataset.

5.4.6 Crowdsourcing quality control

Similar to the Gyms experiment, we defined rejection criteria in advance and categorized jobs as *Red* (rejected) and *Green* (kept for analysis). There was no *Orange* category this time.

Similar to the Gyms experiment, we made sure that participants were able to read a scatterplot. However, since our new scatterplots have many more data points, our screening test was more advanced: participants had to pick the dominant option among five options shown in red, among other unavailable options shown in gray (Figure 5.18). Furthermore, we measured attention with a simple catch question, by asking participants at the end of the job to recall whether the study was about houses and cars, or about other topics (e.g., cameras and cars).

A job was classified as *Red* if the contributor failed the scatterplot test, took an abnormal amount of time to complete the job (1 min or 30 min), or failed the final catch question. Out of the 302 crowdsourced jobs submitted with a valid completion code, 71 (24%) were categorized as *Red*, and 231 (76%) were classified as *Green* and kept for analysis.

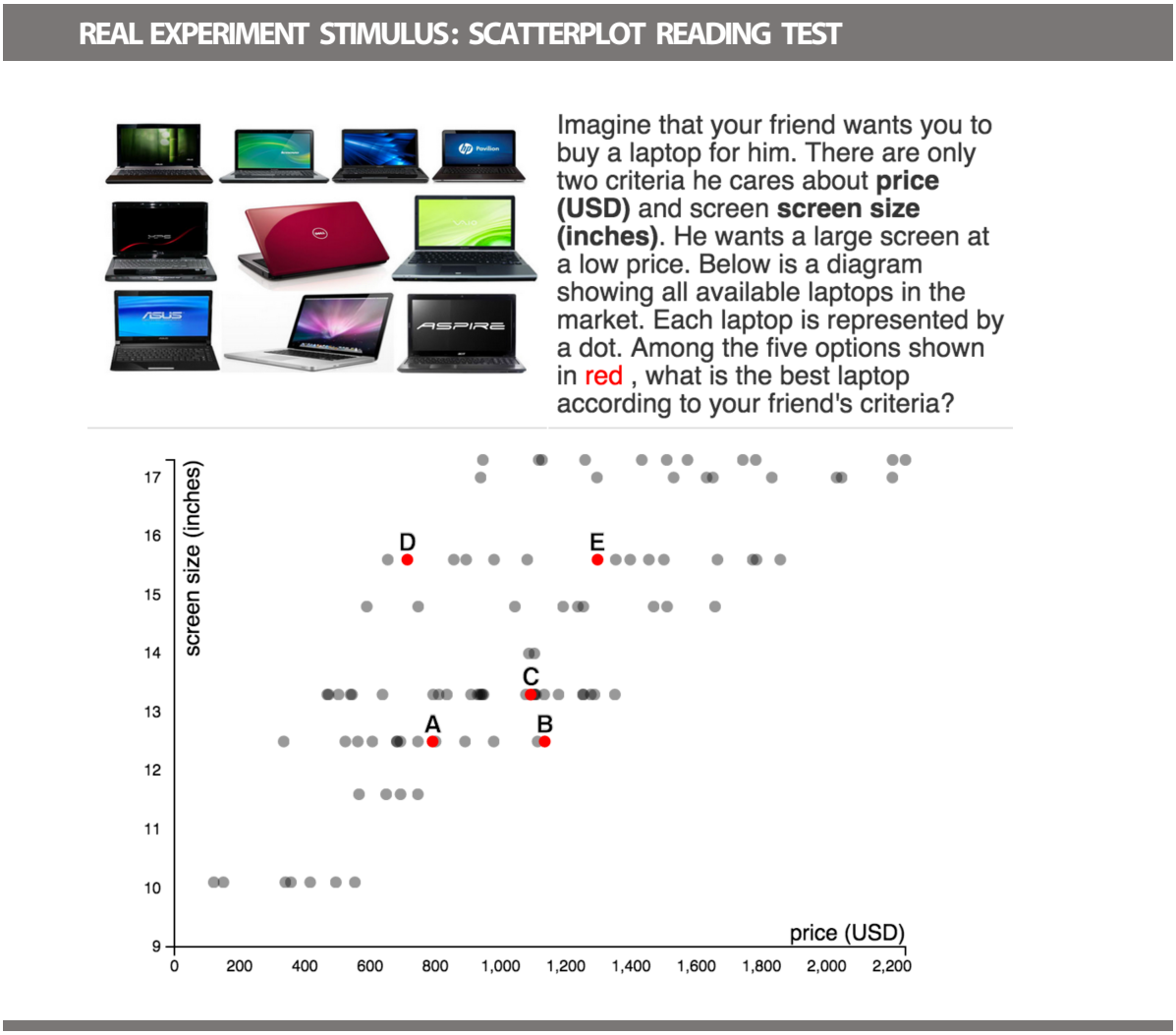


Figure 5.18: The scatterplot reading test. The answer was provided through a radio button.

5.4.7 Experiment design

The experiment followed a mixed design. The independent between-subjects variable was the *decoy position* (*on A* or *on B*), while the independent within-subjects variable was the *dataset* (*houses* or *cars*). Each participant was presented with two choice tasks, one for each dataset. We varied the decoy position within each dataset, resulting in $2 \times 2 = 4$ different pairs of choice tasks. In addition, we counterbalanced the order of appearance of the two datasets, resulting in 8 unique sequences of tasks[3].

As explained in subsection 5.4.5 our dependent variable was the *combined attraction effect*, i.e., the difference between the proportion of participants who chose option A when it was the target, and the proportion of participants who chose option A when the target was B. We compute and report the attraction effect for the *houses* dataset and for the *cars* dataset separately.

[3] houses_A-cars_A, houses_A-cars_B, houses_B-cars_A, houses_B-cars_B, cars_A-houses_A, cars_A-houses_B, cars_B-houses_A, cars_B-houses_B.

5.4.8 Participants

Our study was completed by 231 crowdflower contributors of high quality (level 3) based on their performance on the platform, and whose job was classified *Green*. Their demographics as reported in a post-test questionnaire is summarized in Figure 5.7 (middle map and stacked bar charts labeled "Real"), together with the demographics of the previous experiment "Gyms" and the next experiment "Bets" that we will describe in next section.

As we can see, the demographics between the three experiments are very similar.

5.4.9 Procedure

Pre-test: As explained before and shown on Figure 5.18, participants were first tested on their basic ability to read a scatterplot. Participants who failed this test were removed from the analysis.
Task: Participants then opened a 9-page online form which took on average 5 minutes to complete. For the *house* dataset, they were asked to imagine that they want to buy a house, and that all houses were similar apart from two attributes: size and price. They were then shown a scatterplot with 792 dots representing the houses in the market. Participants were told that they narrowed down their choices to two houses, shown in red. As seen before, the two red dots were labelled A and B, and all other dots where shown in gray. Participants indicated their choice using a separate radio button below the scatterplot. On the next page, they rated their confidence in their choice, justified their choice in a text area, and reported the level of their familiarity with the real estate market. After that (or before, depending on the task ordering), they had to carry out a similar task with a scatterplot displaying 356 cars according to power and efficiency. Participants could review previous pages on the form but not change their answers.
Post-task questions: Participants then had to fill a short questionnaire with their demographic information, and were given the attention test mentioned previously.

5.4.10 Hypothesis

Our statistical hypothesis was that the combined attraction effect will be positive for both datasets.

5.4.11 Results

Experimental stimuli, data and analysis scripts are available at `http://www.aviz.fr/decoy`.

5.4.12 Planned analyses

All analyses reported here were planned before data was collected. Participant choices are shown in Figure 5.8, bars labeled "Real". The top two bars are the proportion of responses for the *house* dataset, with decoys on *price* (A) at the top, and on *size* (B) at the bottom. The next two bars are the responses for the *cars* dataset, with decoys on *efficiency* (A) on the top, and *power* (B) on the bottom. The decoys

are expected to increase the proportion of choices of the target, in the direction indicated by the arrows. As can be seen, this was not the case for either dataset, and the combined attraction effect was even negative in both cases, although small (-3% and -4%).

We now turn to inferential statistics, reported in Figure 5.9 for all our experiments. The two combined attraction effects mentioned previously are shown in pink next to the label "Real". The two dots indicate the point estimates reported previously, and error bars are 95% confidence intervals that indicate the uncertainty around those estimates [123]. Confidence intervals were computed using score intervals for difference of proportions and independent samples. The combined attraction effect was -3%, CI [-14%, +8%] for *houses*, and -4% , CI [-15%, +7%] for *cars*.

Thus, the apparent reversal of the attraction effect observed in Figure 5.8 is way too unreliable for any conclusion to be drawn concerning the direction of the effect [123]. We can only be reasonably confident that the effect is no larger than 15% in either direction. If there is indeed an attraction effect, it is clearly smaller than the combined effect obtained by Malkoc et al. [316] using the classical procedure, and likely smaller than the effect we previously obtained in our Gyms experiment. A "repulsion" effect is also possible. In summary, our results in this experiment are largely inconclusive.

5.4.13 Additional analyses

As can be seen in Figure 5.10, participants reported similar levels of confidence in their answers across both datasets. For the *houses* dataset, participants reported a mean level of confidence of 6.0 on a 7-point Likert scale. For *cars* the mean confidence was 5.9. When it comes to familiarity, responses were diverse for both *houses* and *cars* as can be seen in Figure 5.10, but on average, people were similarly familiar with both datasets (both means 4.2).

5.4.14 Discussion

One major reason for these inconclusive results is a lack of statistical power: since the effect seems small, we would need a remarkably large sample size (i.e., much larger than N=231) or a modified design to be able to reliably assess both the direction and the magnitude of the effect.

Our inability to detect an effect at least indicates that in this experiment, the manipulation we used was not sufficient to trigger the same attraction effect as the effects typically observed in more typical experiments. Attempting to interpret this finding a-posteriori, it may be due to the choice of a constrained choice task. As explained in the background section in section 5.1, the attraction effect requires that participants recognize the dominance relationship between the target and the decoy. Although our participants could read scatterplots (we used a screening test) and the scatterplots we used gave them the opportunity to perceive and to recognize these dominance relationships, it is possible that the dominated alternatives (and thus the decoys) were ignored. Since these alternatives were not part of the choice set, it may have been more clear to participants that they were not needed to carry out the task. In classical attraction effect experiments it is *also* the case that the decoy is irrelevant to the choice task, but participants have to determine its irrelevance (i.e., perform dominance

recognition) themselves. The visual design of the scatterplot (i.e., the two available options shown in red, all other options in gray) may have reinforced the impression that non-available alternatives could be entirely ignored. The disregard of decoys could have also been reinforced by the way participants gave their answers (radio buttons, one for each of the two red choices).

5.5 Bet Experiment: Scatterplot, Many Choices

Given the interpretation discussed in the previous section for the inconclusive findings of our "Real" experiment, it is premature to conclude that the attraction effect does not exist in scatterplots with more than three alternatives. We thus decided to conduct a third experiment named "Bets", where we attempt to improve the previously mentioned limitations in our design. Particularly, we decided to modify the choice procedure (any alternative can be selected), the datasets (synthetically generated to maximize the effect), and the experiment design (within-subjects to maximize statistical power).

5.5.1 Design rationale

Here we describe and motivate the design of this new experiment.

5.5.1.1 Replicated study

Most attraction effect studies (including our previous experiments) follow a between-subjects design. However, these designs typically suffer from low statistical power. The width of confidence intervals in our previous experiments indicates this was the case there.

We therefore decided to adopt a within-subjects design. Wedell [500] was able to measure a clear attraction effect with numerical tables using a within-subjects procedure, where participants were given multiple choice tasks. He further tried to increase statistical power by *i)* excluding no-decoy conditions and only measuring the combined decoy effect, and *ii)* choosing a scenario with which people were less familiar (lotteries) in an attempt to amplify the effect [336, 392], since in our a-posteriori question on familiarity in our first replication experiment (Figure 5.10) indicated that our participants were very familiar with gyms.

5.5.1.2 Scenario and attribute values

Wedell's scenario involved choosing among three lottery tickets, each defined by two attributes: the probability of winning (*probability*), and the amount that can be won (*prize*). Participants were presented with twenty choice tasks in sequence. Each time, three lottery tickets were presented and participants had to choose one. Wedell thought that the abstract nature of the task and of the attributes would reduce possible carry-over effects, such as participants building up strategies based on past choices.

The non-dominated alternatives (targets and competitors) used in all Wedell's tasks were taken from a pool of five alternatives (A to E in Figure 5.19). All had the same expected value of $\sim$\$10. Thus,

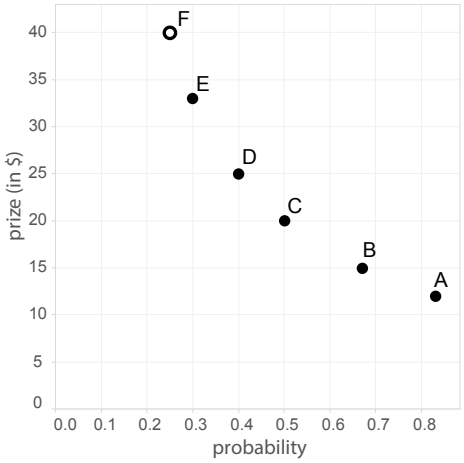


Figure 5.19: The non-dominated alternatives used in our choice tasks.

	A	B	C	D	E	(F)
probability	0.83	0.67	0.5	0.4	0.3	0.25
prize	$12	$15	$20	$25	$33	$40

though a rational choice maker would only need to compare alternatives along a single dimension (expected value), the choice tasks had the same dominance structure as tasks involving two independent attributes such as in the previous gym experiment.

For each possible pair of alternatives in (A,B,C,D,E) Wedell generated two choice tasks, one with a decoy on probability, and one with a decoy on prize. We use the notation XY to refer to a task where X is the target and Y is the competitor, and refer to the two choice tasks XY and YX as *matched*. For example, the pair of alternatives (A,C) yields the two matched tasks AC (where the decoy is on A) and CA (where the decoy is on C). Wedell's design resulted in 10 pairs of matched choice tasks (20 tasks in total).

Although we planned to reuse the same targets and competitors, it appeared that the distance between the target and the competitor was visually very small in some scatterplots compared to others. Thus we added an alternative with the same expected value (F in Figure 5.19) and excluded all tasks that involved adjacent target/competitor pairs (e.g., AB, or DE). This new design also resulted in 10 pairs of matched choice tasks, and 20 tasks in total.

5.5.1.3 Adding distractors and decoys

While Wedell only added one decoy to each of the choice tasks, our goal was to present many alternatives as explained in the previous section. For each pair of matched choice tasks, the procedure consisted of two steps. We explain the procedure for AC and CA (see results in Figure 5.20), but it is the same for all other pairs:

Step 1. A baseline choice task analogous to T_0 in Figure 5.11 was created by adding distractors

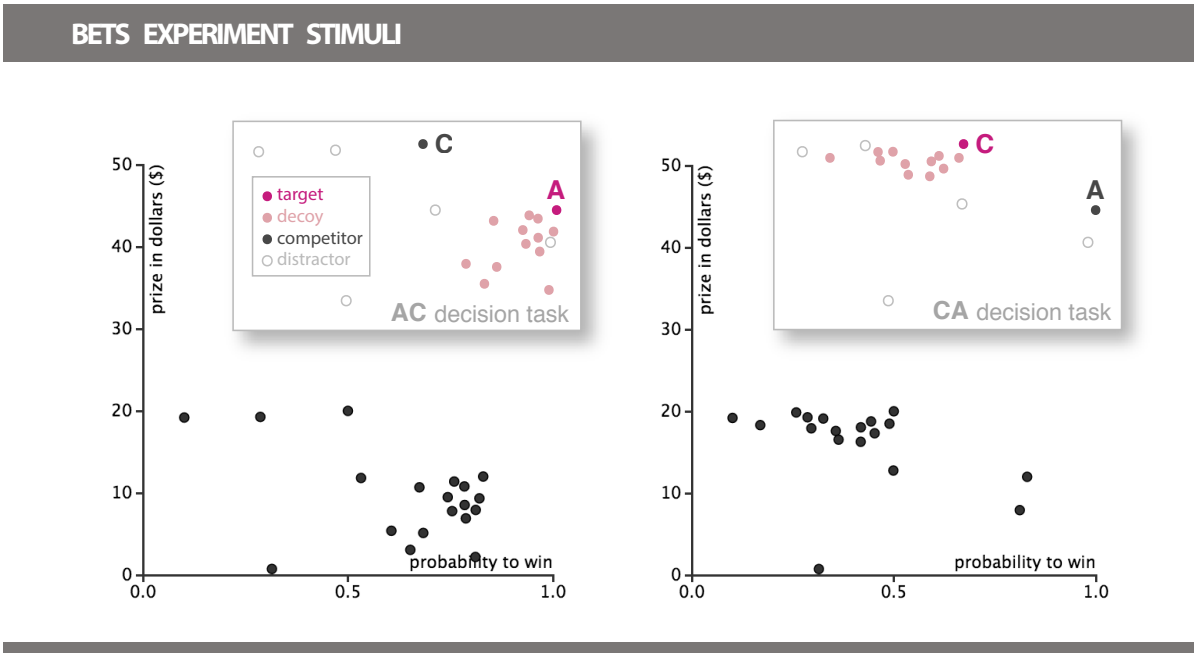


Figure 5.20: Experimental stimuli for the two matched choice tasks AC and CA (black-and-white background images), and explanatory annotations (box overlays). See Section 5.3 for the full details.

dominated by A and/or C. One or two distractors (number randomly drawn) were added in each of the regions R_A, R_C and R_{AC}, following a uniform spatial distribution.

Step 2. Two separate choice tasks AC and CA were then created by adding decoys as shown in Figure 5.11. For the task AC (decoys on A), 10 to 20 decoys (number randomly drawn) were added to the region R_A following a bivariate half-normal probability distribution. On each axis, the mode of the half-normal was A's value on this axis, and the mean was this value multiplied by 0.7. The use of half-normals yielded decoys that tend to cluster near A, but whose density smoothly decreases with distance to A for a more natural look. The same was done for the choice task CA.

In both steps, overlaps were eliminated by *i)* defining overlap between two alternatives as a distance less than 0.025 in normalized coordinates (prize divided by 40, probability left unchanged) and *ii)* whenever a new alternative is randomly drawn, iterating until there is no overlap. The reason why the number of alternatives to draw was randomized (i.e., 1–2 for each region in Step 1 and 10–20 in Step 2) was to create more variation across scatterplots and make it more difficult for participants to infer patterns in the experiment.

5.5.1.4 Ordering of choice tasks

Wedell [500] felt that the abstract nature of the task could help participants forget the values across repetitions. As the attraction effect he measured was robust and consistent with the placement of the decoy, his design seems justified. We followed the same design. Our presentation order for the 20 choice tasks according to Wedell [500], but modified to account for our different set of tasks and for the

fact that we present each task on a separate Web page, while Wedell used a four-page paper-and-pencil test.

We created a task ordering such that *i)* a choice task and its matched task (e.g., AC and CA) are always at least 5 pages apart; and *ii)* the role of an alternative alternates over time. For example, if D appears as a target in a task, it will be a competitor the next time it appears. To reduce further possible ordering effects, we created a second ordering where each task is replaced with its matched task. Participants were randomly assigned to each ordering. [4]

To make it more difficult for participants to infer patterns in the sequence of choice tasks, we additionally inserted seven irrelevant choice tasks at various positions, which were not used in our analyses. These tasks differed in that they had either one or three non-dominated alternatives (instead of two), and they did not exhibit an imbalance in the number of asymmetrically dominated alternatives.

To win a good bonus, start your job by deleting all "bad" tickets first.

5.5.1.5 Stimuli: interactive scatterplots

In this experiment, we added minimal interaction to the scatterplot visualizations. In the previous experiments, the scatterplots were static and each data point was labeled with a letter (Figure 5.6), so that participants could specify their choice through separate radio buttons. As we are now dealing with more data points, labels were removed to prevent clutter (Figure 5.20), and participants were asked to specify their choice by selecting the data point. Points were highlighted when hovered. Hovering a point also displayed horizontal and vertical projection lines, and the data point's X and Y values were overlaid on the axes. Such interactions help examine the data and are not uncommon in scatterplot visualizations. After a point was clicked, its color changed and the participant was asked to confirm her choice by clicking on a button at the bottom of the page.

We added a short flicker during task transitions in order to elicit change blindness and prevent participants from easily detecting similarities and differences between two successive scatterplots.

5.5.1.6 Crowdsourcing quality control

We made two major modifications to the previous procedures: *i)* we added a preliminary tutorial, *ii)* we used a real decision making task where choices affected subsequent monetary gains.

The tutorial simultaneously explained the scenario (the lottery tickets, and what their probability and prize meant), and how to read scatterplots. Although Wedell [500] did not provide similar training, crowdsourced contributors do not necessarily have the same qualifications as university students, and the notion of probability in particular is known to be challenging [332]. In order to prime participants to use their intuition rather than doing calculations, probability was explained qualitatively rather than quantitatively.

[4] The two orders were: [CE BD AC FA DF FB CF AD EC BE DB EA FD AF CA FC BF AE EB DA] and [EC DB CA AF FD BF FC DA CE EB BD AE DF FA AC CF FB EA BE AD].

After the tutorial, participants were given a test question consisting of choosing one among 13 lottery tickets presented as a scatterplot. Three tickets were non-dominated (and thus formally uncomparable), and the remaining 10 were considered wrong answers.

In order to better approximate real-life decisions and motivate our participants, we informed them that a computer will run the lottery after the experiment is completed, and for every winning ticket they picked, they will be payed a bonus proportional to the ticket's prize (more details for the bonus calculation in section 5.5.2.1). The use of a real decision task with consequences is common in behavioral economics and is occasionally used when studying the attraction effect (e.g., choosing between objects or money [435]).

Similarly to our previous experiment, we defined our rejection criteria in advance and categorized jobs as *Red* (rejected and not payed), *Orange* (payed but not analyzed) and *Green* (analyzed). A total of 120 jobs were submitted with a valid completion code. A job was marked as *Red* (12%) if its completion time was abnormal (0.8%), if the contributor failed the tutorial test (11%), or if during the experimental trials, the contributor selected a dominated option more than half of the time (12%). A job was marked as *Orange* (27%) if the contributor always chose the highest probability (27%) or the highest prize (0%). These contributors had a too strong prior preference (in this case, risk aversion) to be sensitive to the attraction effect. Their answers will not change the overall sign of attraction score (given the symmetry of the design), but by removing these answers from the analysis allowed us to increase statistical power. The remaining 61% (N=73) were marked as *Green*.

5.5.1.7 Reasoning over an attraction effect pattern

As discussed in subsection 5.1.1, there are many theories about how people perceive the relation between target and decoy. Especially in the scatterplot case, we do not know how people could interpret the existence of multiple decoys below the target. Thus, at the end of the study, we displayed one of the study scatterplots with label A on the target and posed three post questions regarding the existence of multiple points below A: to report how much it affected their choices during the study, to give a rational explanation for their existence, and whether they came up with it on the spot or had it in mind during the study.

5.5.2 Experiment design

The design consisted of two within-subjects factors: *task pair* (10 pairs of matched tasks), and *decoy position* (on *probability* or *prize*).

5.5.2.1 Procedure

Briefing: We first briefed our crowdsourced contributors that they will have to choose lottery tickets and will receive a bonus for each winning ticket, for a total of $0.60 on average. We explained that the task lasts about 15 min, does not require prior math knowledge, but that they would have to

make judgments based on simple diagrams. We noted they would be provided with a quick tutorial on reading diagrams, and they would have to pass a related test question successfully to continue. We further explained that we would run the lotteries at a later time, and that they would receive their final bonus within 1 to 3 days. We clarified that the virtual dollars seen during the study are proportional to real dollars ($0.10 per $1.97).

Training and Pre-test: They then opened an external link to the 10-page tutorial. The tutorial presented the probabilities in the context of lottery tickets. We first explained the notions of prize as the amount of money you get if the ticket is a winning ticket, and the probability that the ticket is indeed a winning ticket. We gave simple numerical examples of a ticket with $10 prize and its probability as a number between 0 and 1. We explained that if the ticket has probability 0, they will certainly lose, and if the ticket has probability 1, they are guaranteed to win. In all other cases, they do not know if they will win or not, but the higher the probability the more likely it is. Then we moved on to a textual numerical example that compared tickets with the same prize and different probabilities, explaining that the optimal choice is the one with higher probability. Next, we showed the same example in a scatterplot representation. Similarly, we showed other scatterplots with two tickets with same probabilities and different prizes, and two tickets where one was dominant in both attributes. We later showed a 7-point scatterplot with only one dominant ticket in both attributes. Finally, we showed a 2-point scatterplot with formally incomparable tickets where there is no good or bad choice. We advised them to follow their intuition to decide in such situations. Before the end of the tutorial, we informed them that for the rest of the job they will deal with real gambles and that they have to start with a test task. The test task consisted of a 13-point scatterplot with three incomparable dominant choices together with multiple dominated ones, and we asked them to choose one. If participants chose any of the three they could proceed to the actual experiment. Depending on their answer (correct or wrong) each participant was given a different training completion code, that they had to paste back in the platform.

Pre-test: Motivation: Contributors who chose a valid ticket on the test were told that the ticket won, and that they would get a $0.10 bonus for the ticket if they proceed and complete the job. We encouraged them to continue with the job to win more tickets. If their choice was incorrect, we advised them to quit the job since we would have to reject it –but did not prohibit them from continuing.

Main Study Task: Participants then opened a second external link to the main study, a 31-page form, where they saw the twenty choice tasks, mixed with the seven distractor tasks. In each scatterplot task they clicked on the lottery ticket of their choice and progressed to the next choice task.

Post-test: After completing all choice tasks, participants rated their overall confidence, their perceived difficulty of the job, their familiarity with gambling games, and whether they knew of the notion of "expected value" in probability. They then filled a short demographic questionnaire and received a unique study completion code that they pasted onto crowdflower. Participants finished their job, by responding to effect interpretation questions (described in section 5.5.1.7).

Debriefing: All participants received a baseline payment of $0.20, while *Orange* and *Green* received a

bonus of $0.10 plus a lottery bonus. The expected lottery bonus was $0.50 if no dominated alternative was chosen, based on a conversion rate of 0.0025 between the scenario's "virtual dollars" and USD. After the experiment was over, we determined each lottery bonus by *i)* running Bernouilli random draws to determine the winning status of each chosen ticket, *ii)* summing up the prizes of winning tickets *iii)* multiplying by the conversion rate. These bonuses were then paid.

5.5.2.2 Participants

Our participants were 73 crowdflower contributors whose job was marked *Green*. Their demographics, shown in Figure 5.7, were similar to the first and second experiment.

5.5.2.3 Hypothesis

Our research hypothesis was:

Hr3 When a choice task with multiple alternatives is presented as a scatterplot, the addition of decoys increases the attractiveness of the target.

This translates into the following statistical hypothesis:

H3 The mean *attraction score* will be strictly positive (this metric will be explained in the Planned Analyses section).

5.5.3 Results

5.5.3.1 Planned analyses

We first report descriptive statistics of participant choices in a similar way to Wedell [500]. We recorded a total of 1460 choices (73 participants $\times$ 20 choice tasks). We pair choices according to matched tasks (e.g., tasks AC and CA in Figure 5.20), yielding 73 $\times$ 10 = 730 choice pairs. Of all these choice pairs, only 24 (3.3%) included a dominated alternative. Wedell reports similar results (2%), even though his tasks only involved a single dominated alternative.

Figure 5.21 summarizes the remaining 706 choice pairs as a contingency table as we see in the bottom right cell of the table of Figure 5.21, shown next to Wedell's on the left (730 pairs minus 3% of decoy selections). Contingency tables present the relationship of two (or more) categorical variables. Our two categorical variables are: the participant choices when the decoys favored probability and their choices when the decoys favored prize. The choice "prob" indicates that participants choose the bet with the higher probability and, "prize" the bet with the higher prize. Choice pairs fall into four categories. One is choosing the ticket with highest probability in both tasks (i.e., ticket A in Figure 5.20). This represents 59% of all choice pairs, and is reported in the top-left cell in Figure 5.21. A second possibility is choosing the ticket with highest prize twice, which represents 10% of all cases. The remaining two possibilities, shown in bold cells, consist in always choosing the target (23%), or always

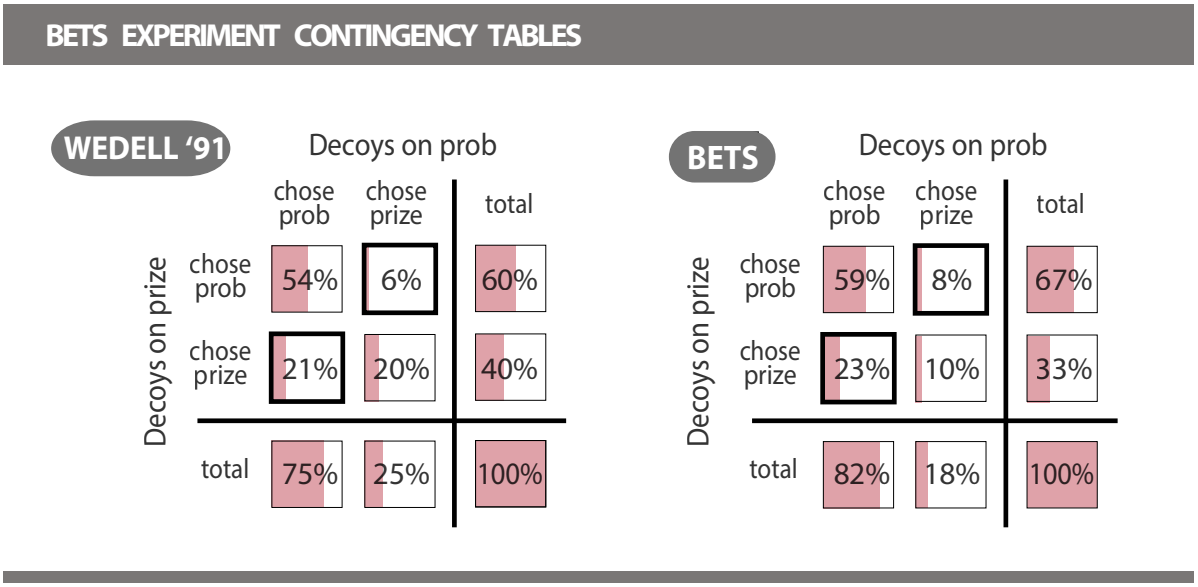


Figure 5.21: Contingency tables showing choice pairs for all matched tasks

choosing the competitor (8%). All perimetric cells are the additions of their respective columns and rows.

The patterns in our contingency table follow Wedell's closely [500]: participants favoured higher probability overall (reflecting again risk aversion), but when their choice was inconsistent across two matched tasks, they chose the targets more often than they chose the competitors. We now turn to inferential statistics.

Similarly to Wedell, we used as dependent variable an *attraction score*, calculated on a per-participant basis as follows. Each of the 20 choice tasks was assigned a score of 1 when the ticket with highest probability was chosen, a score of 0 when the ticket with highest prize was chosen, and a score of 0.5 when another (dominated) ticket was chosen. Then, we averaged all scores for the 10 choice tasks where the decoys were on probability (yielding a score S_{prob}) and did the same for the 10 tasks where the decoys were on prize (yielding a score S_{prize}). The difference between the two scores $S\ S_{\text{prob}} - S_{\text{prize}}$ was the *attraction score*.

A participant who is not subject to the attraction effect should exhibit the same preference for high probability irrespective of the position of the decoys, thus her attraction score should be close to zero. We multiplied the attraction score by 100 to obtain a percentage analogous to the combined decoy effect reported in the gym experiment. The difference here is that the percent difference is computed within-subjects instead of between-subjects, and it incorporates choices of dominated options as "neutral" observations.

The mean attraction score was 15%, with a 95% bootstrap confidence interval of [10%, 22%] (see Figure 5.9). Thus we have very solid evidence for **H3**, even if the effect is smaller than in Malkoc's gym study [316]. We cannot directly compare our effects with Wedell's [500] due to the use of different

statistical methods, but Figure 5.21 suggests the effect sizes are comparable.

5.5.3.2 Additional analyses

As shown in Figure 5.10, participants reported various levels of familiarity with gambling and were confident in their choice overall, although slightly less than in the gym experiment. Data on participants' knowledge of expected values was missing due to a bug.

Concerning the final questionnaire on how participants interpreted the presence of decoys (see Section 5.5.2.1), 8 participants reported not being able to see the scatterplot image, leaving data from 65 participants. When asked whether the uneven distribution of tickets affected their choices, 41% replied "never" or "rarely", 46% replied "sometimes", 12% replied "often", and none replied "always". When asked why they thought there were more tickets in one region than the other, most (86%) gave responses that were irrelevant or unintelligible based on an informal content analysis of open text responses. Out of the 9 remaining responses, 5 referred to a strategy employed by the lottery organizer (e.g., *"To tempt people to choose tickets of high prize but with low probability, increasing the profitability of lottery owner"*; *"To distract from choosing the higher chances of winning"*), and 4 referred to tickets as past choices from other players (e.g., *"Customers want to win a higher prize"*; *"Maybe more people played the same"*). Only 4 participants (quoted here) reported that they had their explanation in mind while performing the task, while the other 5 reported that it was prompted by our question. Thus there is little evidence that participants' preference for the target was motivated by deliberate, reasoned strategies.

5.6 Discussion

The Bets experiment confirms that an attraction effect can be observed in choice tasks involving more alternatives than simply a target, a competitor, and a decoy. Bettman et al [59] expected that the effect would disappear as more alternatives are added, since pairwise comparisons and dominance recognition becomes hard if numerical tables are used. Our experiment indicates that this may not be the case when using visualizations, as visualizations such as scatterplots support fast comparisons and dominance recognition.

Despite the inconclusive "Real" experiment, the Gym and Bets experiments suggest that the attraction effect generalizes to data visualizations, and thus contribute to the ongoing debate in decision-making research on whether the effect can generalize to non-numerical formats [174, 241, 434, 523].

There are several potential limitations to our study. One is that the presence of an attraction effect is not necessarily a proof of irrationality, as the way dominated alternatives are distributed can in some cases provide relevant information. For example, a real estate investor may infer from a region with many dominated alternatives that a certain type of house is more common, and therefore represents a larger market. However, situations also exist where the number and position of dominated alternatives is clearly irrelevant and where a preference for the target would be irrational. This was the case for

our bet experiment, and our data does indicate that the vast majority of our participants were unable to rationalize their choices based on where the dominated alternatives were located.

Another limitation is that we tested very specific datasets, i.e., synthetically-generated datasets with only two non-dominated options and a large number of decoys. More realistic datasets need to be tested, although our inconclusive results with real datasets suggest that the effects may be harder to measure [133]. Also, as discussed in subsection 5.3.2, testing more than two non-dominated alternatives would depart from a "pure" attraction effect experiment.

Finally, since in the bet experiment regions with many decoys were visually more salient, it is possible that they drew participants' attention towards the target, or similarly, that participants sometimes failed to see the competitor because it was an isolated point. This explanation is not incompatible with the existence of an attraction effect, but it does suggest that part of the effect with scatterplots (but not with numerical tables) may have perceptual origins.

5.7 Conclusion

This chapter examined whether participants can be subject to a cognitive bias while using a visualization. In particular, the chapter investigated whether a decision can be "correct" (choosing dominant data-points), yet irrational, since participants' dominant choices can be influenced by factors irrelevant to the decision.

The work presented in this chapter was the first study of the attraction effect in visualizations in a series of three experiments:'Gyms", "Real" and "Bets". Despite the inconclusive results of the "Real" experiment, the experiments "Gyms" and "Bets" suggest that the attraction effect generalizes to data visualizations. While the first experiment focuses on a traditional procedure with only two or three alternatives, the"Bets" experiment shows that the effect can persist with more alternatives. Bettman et al. [59] expected that the effect would disappear as more alternatives are added, since pairwise comparisons and dominance recognition becomes hard if numerical tables are used. The findings suggest that this may not be the case when using visualizations, as visualizations such as scatterplots support fast comparisons and dominance recognition. Overall, the attraction effect study indicates that when people visualize choice alternatives using scatterplots, the number and position of irrelevant (dominated) alternatives may influence their choice. This shift in preferences violates basic axioms of rational choice theory [240]. In addition to being the first InfoVis study on the attraction effect, the work presented in this chapter contributes to the ongoing debate in decision-making research on whether the effect generalizes to non-numerical formats [174, 241, 434, 523].

The takeaway message of this work is that cognitive biases can affect decisions even if the data is well visualized and fully understood, thus traditional visualization design rules may not apply when the goal is to support decision making. Therefore, the following chapter will investigate alternative ways to alleviate the attraction effect in visualization systems that target decision support.

Chapter 6

Towards Improving Decision Support Visualization Systems

This chapter explores two approaches for improving visualization systems in a way that can help people make better decisions. First, Section 6.1 will revisit the visualization systems that target decision support (reviewed in the Background section 2.4). Taking into account the empirical findings of previous chapters, this chapter introduces a new scatterplot-based decision support tool, named DcPairs. A use-case scenario of a prospective undergraduate student choosing a university from the "QS world university ranking" dataset illustrates the functionality of the tool.

Chapter 5 showed that visualization systems can be prone to biases, such as the attraction effect, where inferior and irrelevant data can affect users' choices. As DcPairs is not meant to address cognitive biases, it likely suffers from similar issues. Therefore, the rest of this chapter will consider debiasing techniques. Section 6.2 will first review the current approaches in cognitive bias alleviation. Section 6.3 will follow a design-based approach and will investigate a novel debiasing technique inspired by the decision strategy "elimination by aspects (EBA)".

6.1 DcPairs: Multi-attribute Decision Support with Annotations

This section is based on a collaboration with Paola Valdivia and Christoph Kinkeldey for a poster article in EuroVis Conference [136] and will present DcPairs, a visualization system that targets decision support. Each of the following sections will discuss certain design choices that led to the development of DcPairs.

6.1.1 The use of color in decision support visualizations

Visual variables often impede judgments in unexpected ways. For example, in Chapter 5 the **position** of inferior data points influenced the perceived attractiveness of other data points. Similar cognitive

biases exist (previously discussed in section 2.3) in which a data point is perceived as more attractive, if positioned in the middle [433], or closer to a better but unavailable one [376]. Moreover, the first arithmetic **value** people see can anchor their probability estimations [181] (section 2.3). People's ability to recall can also be affected by whether the **shape** of an item is perceived as bizarre [327] or distinct among a list of similar others [368] (section 2.3). Visualization research has studied extensively the perceptual properties of visual variables (e.g., position, size, shape, color, orientation, texture) [105], but the cognitive properties of these variables have received less attention.

Color hue is considered as the second best effective variable for categorical data after the spatial position [347]. Despite the perceptual benefits of using color, its cognitive properties have not been much investigated in the context of visualizations. According to the current list of cognitive biases (Appendix A), there does not seem to be a study indicating a cognitive bias towards a particular color (e.g., people always choose a blue item). But most colors are associated with certain semantics. For instance, red is often associated with "danger" and green with "life" [495]. Such semantics can differ between cultures in a way that it can be hard for a designer to foresee. For instance, in the Chinese culture, red is seen as a metaphor for "life" and green for "death" [495]. Also, a recent visualization study showed that some color palettes communicate different emotions to users [45] (Figure 6.1). Along these lines, another study in business dashboards suggests that our brain tends to assign meanings to distinct colors and that when multiple colors are used (e.g., in a bar-chart) decision makers can be distracted and need more time to come to a decision [55]. For this reason, in decision making, where subjectivity and context play an important role, the use of color should be handled with care.

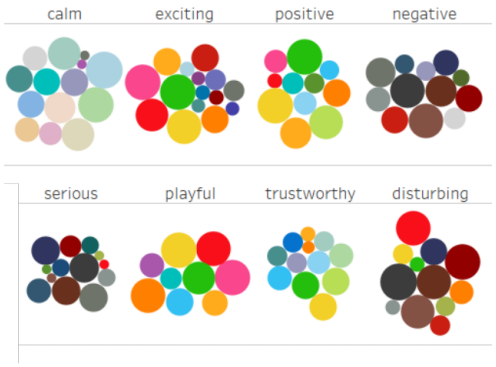


Figure 6.1: Bartram et al. [45]

As illustrated in Figure 6.2 and previously noted in Chapter 2.4, most visualization systems that target decision support choose to encode the identity of attributes with a distinct color to help identify them across multiple views [35, 85, 201, 364]. Systems such as WeightLifter [364] (Figure 6.2 a), LineUp [201] (Figure 6.2 b), Value Charts [51], LiteVis [439], or STRATOS [35] (Figure 6.2 c) utilize color to encode each attribute identity in stacked bar, tabular views, or parallel coordinate views. Other tools make use of color to encode different information. For example, the Data Context Map tool [93] highlights the result sets of different user queries in color, so it becomes evident which result sets belong to which query and where the sets overlap. The AHP Treemaps tool [33] maps each choice alternative to a different color while presenting them in multiple hierarchical layers (Figure 6.2 e).

Using colors to label attributes or alternatives is a straightforward way of representing variables across multiple views (known as "nominal information coding" [495]). However, this approach has three main limitations. First, as mentioned above, colors may influence the perceived meaning and importance of the variables they represent. For example, in a car purchase, if the horsepower attribute

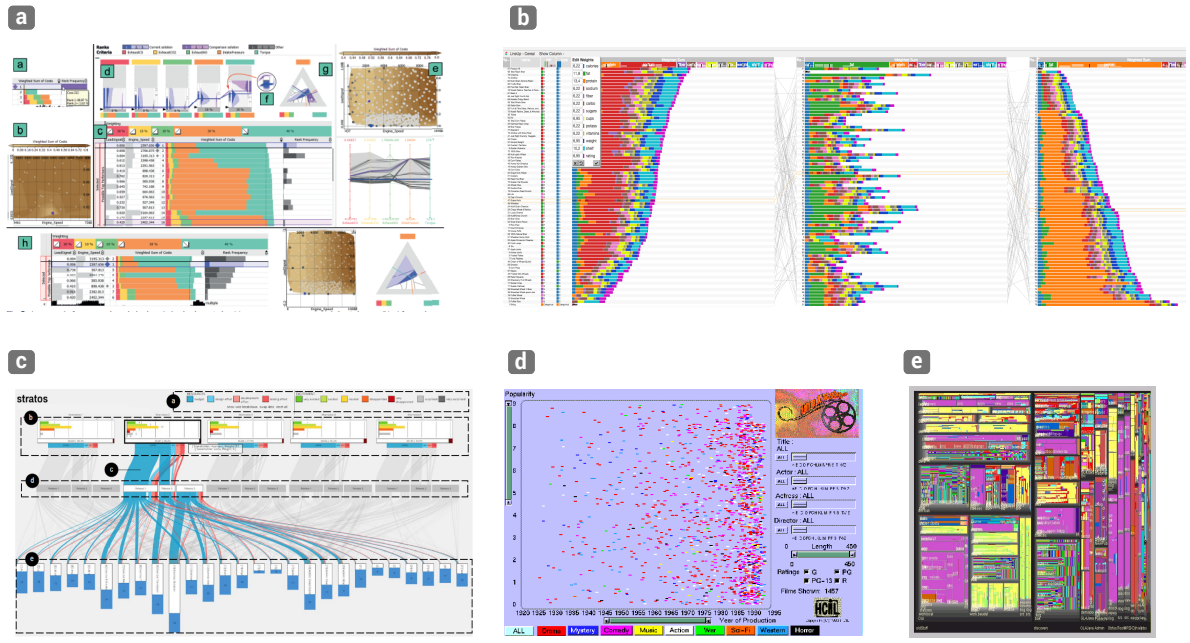


Figure 6.2: This Figure illustrates the use of color in current visualization systems that target decision support. Most systems use color to encode the identity of attributes such as WeightLifter (a [364]), LineUp (b a screenshot from the online version of the tool using the cereals dataset [10, 201]), STRATOS (c [35]). And others (e [13, 33]), to encode choice alternatives such as FilmFinder (d [16]) and AHPtreemaps.

is marked with red color, could this affect its perceived importance over the number of the airbags attribute? The second limitation is scalability. For visual interfaces, a maximum number of 6-12 distinct colors has been suggested [347, 495], which can be limiting for a common multi-attribute choice task. Moreover, the upper limit of 12 colors does not necessarily translate to a display of 12 attributes, because the number of colors must account for the background and default object colors, even when they are black, white or gray [347]. The list of available colors can be limited further if we consider color deficiency since about 1 out of 10 men can not distinguish colors that differ in the green-red direction [495]. For instance, ColorBrewer recommends color blind safe qualitative color schemes with up to 4 colors only[3]. Third, color is an overloaded visual variable often needed to encode critical information. For example, LineUp uses red and green to encode attributes, but also to indicate which alternatives moved up or down when the ranking changes [201]. WeightLifter re-uses color purple to highlight the chosen alternative and the blue to highlight the optimal one [364].

The design of DcPairs considers an alternative approach that does not encode decision attributes by color, in contrast to most decision support visualizations systems.

6.1.2 Attributes encoding in decision support

Usually, it is the representation of data cases that receives the most attention during the design of a visualization system. In the context of a multi-attribute choice task though, the representation of attributes can be equally or even more challenging. Most systems employ multiple views for manipulating attributes. For example, WeightLifter [364] in Figure 6.2 a offers one view for adjusting attribute weights (triangular widget), another view suggesting the optimal alternatives (tabular stacked bars), one for attribute value retrievals (parallel coordinates), and, finally, a scatterplot for certain overview tasks. To coordinate the views, WeightLifter utilizes brushing and linking for the alternatives and distinct colors for the attributes. Using color seems to be the main strategy in decision support visualizations to link the attributes along with the choice alternatives.

Mapping both attributes and alternatives across multiple coordinated views can be challenging for a designer and, as shown in Figure 6.2, it can lead to designs overloaded by color. Since the approach in this section is to avoid color encoding for attributes and alternatives, the DcPairs system will have only a single view rather than multiple coordinated views.

6.1.3 The use of scatterplot matrices in decision support

As discussed in Chapter 2.4, most decision support visualizations base their design on tabular visualizations [35, 85, 201, 364]. The results in Chapter 3 indicate that tabular visualization has indeed a speed advantage in comparison to scatterplot matrices and parallel coordinates. A slight disadvantage of tabular visualizations is that most designs offer a limited overview support, requiring the addition of extra views (e.g., stacked bars, scatterplots, parallel coordinates [201, 364]). On the other hand, scatterplot matrices showed an advantage for overview tasks in Chapter 3. Participants were significantly faster with scatterplot matrices in correlation tasks, and they reported to trust more choices made with it. A possible explanation is that scatterplot matrices support overview tasks (confirmed by the results with the correlation task), making participants more confident that they did not miss a particularly interesting alternative. Moreover, in WeightLifter's evaluation [364], expert analysts requested the addition of a scatterplot for overview tasks. Considering also that scatterplot matrices and tabular visualizations achieved similar decision accuracy and user preference for decision tasks, the scatterplot matrix could be a viable alternative, which in addition can reduce the need for multiple views due to its overview advantage.

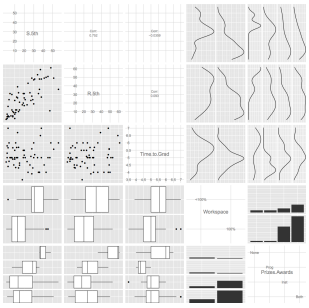


Figure 6.3: A Generalized Pairs Plot [157]

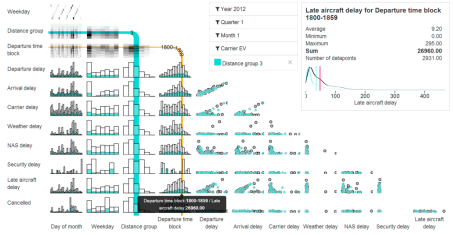


Figure 6.4: GPLOM [246]

An important concern with a scatterplot matrix is that it can be only used with quantitative attributes. Nevertheless, Emerson et al [157] defined the **generalized pairs plot** (shown in Figure 6.3) , which we will refer to as **pairs plot** for simplicity, as a grid of plots displaying both quantitative and categorical attributes. Scatterplot matrix has since been

used for quantitative attributes [104] by adding within the matrix other types of representations such as mosaic plots [211], for categorical or side-by-side boxplots [469] for qualitative-categorical attribute pairs.

Pairs plots offer a systematic way to compare attribute pairs in a single compact view [347]. Due to its symmetric nature, half of the matrix as well as the diagonal are usually omitted [246, 347] providing additional design space to be filled with histograms [359], color coded interactive menus [126] or other interactive charts [246] (see GPLOM in Figure 6.4). In decision making, pairs plots can be particularly useful to find trade-offs between pairs of attributes (e.g., quality vs. price).

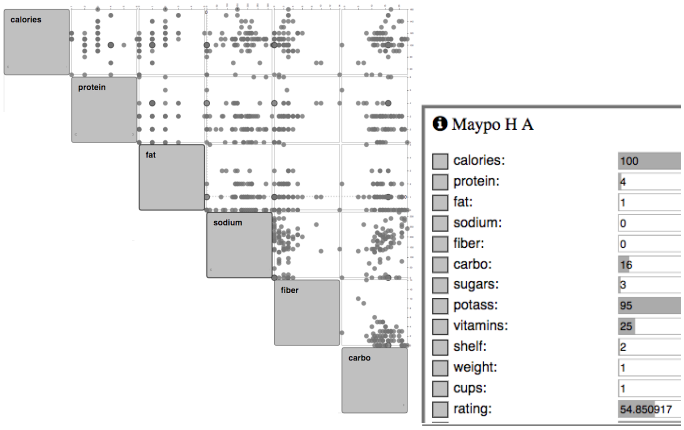


Figure 6.5: Scatterplot Matrix DcPairs

The scatterplot matrix (a type of pairs plot), is deemed to scale up to a dozen attributes and about a hundred alternatives, and is known to be especially suited to highlight attribute *relations*, *trends* or *outliers* [347]. For example, the *relation* of a quality index and the price of a product could help identify good deals; many alternatives of similar attributes could indicate market popularity (trend); consumers could have predefined attributes preferences (e.g., brand), but still be interested in seeing an extremely cheap offers that do not necessarily fitting their criteria (outlier).

For these reasons, the design of DcPairs system is based on a pairs plot that consists of square-shaped elements showing scatterplots for pairwise comparison of attributes (Figure 6.5). Since the matrix is symmetric, only the upper half is shown. Once the user hovers over a dot, an inspector window (bottom right) is triggered to retrieve the values for all attributes. Even though pairs plots, and scatterplot matrices in particular, have been used extensively in visual analysis and exploration [156], they are less widespread in decision-support tools.

6.1.4 The use of personal annotations in visualizations

As mentioned in Chapter 2.4, a limitation of most visualization systems that target decision support, is that they do not aid users who want to enter new data or metadata into the system. For example, LineUp [201] is illustrated through a scenario of a prospective student who wants to choose a university using a dataset with attributes, such as academic reputation, citations or faculty-student ratio. The admission fee is not included in the universities dataset, but it can be a very important attribute for a student on a limited budget. However, the student may not have access to data about all university fees. So, as soon as she converges to a small subset of universities that satisfy other criteria (reputation, citations, etc.), she may further refine her choices using fee information that she finds online. Taking into account external knowledge in multi-attribute choice tasks can be useful in decision making.

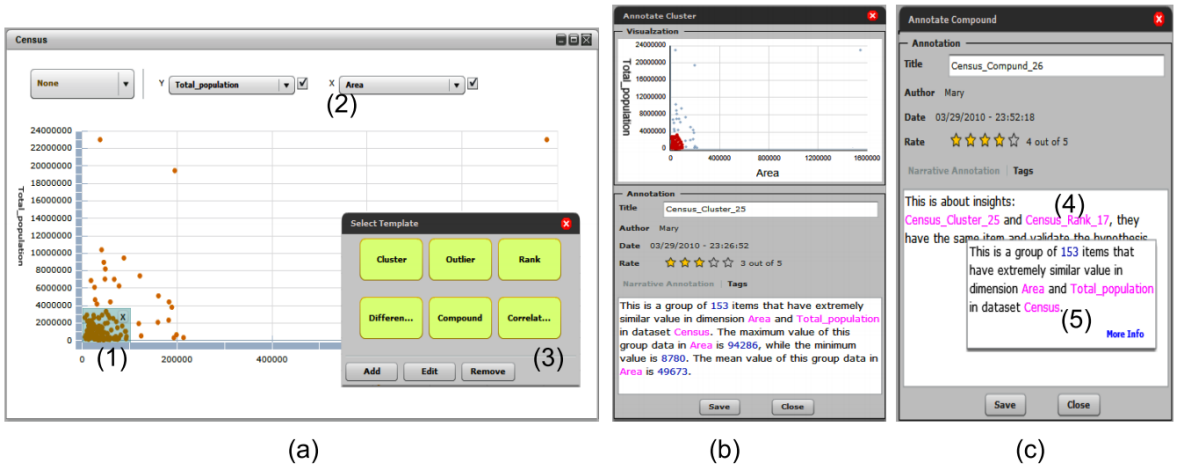


Figure 1: Semi-automatic annotation generation using Click2Annotate. (a) A scatterplot with a cluster in it and the annotation process. (b) The automatically generated annotation for the cluster. (c) The annotation generated for a compound fact.

Figure 6.6: Figure from Click2Annotate prototype by Chen et al. [91]

Therefore, the design of DcPairs will also provide means to support data input from the user.

Generally, annotations are used to provide metadata for any kind of datum to explain it further, comment on it, etc. In the context of visualizations, the annotate analytic task consists of the addition of attributes by the user in the form of "*graphical or textual annotations associated with one or more visualization elements*" [70]. As illustrated in Figure 6.6, annotations are often suggested in supporting insight externalization [91]. Zhao et al. also showed how user-authored annotations could effectively support the generation of hypotheses and the activity of sense-making during exploratory analysis of graphs [532]. Moreover, in a business intelligence analysis scenario, Elias, and Bezerianos [153] gathered insights on the use of annotations from expert interviews to find that annotations can successfully support visual analysis in multi-chart visualizations, e.g., for linking and organizing the charts during the analysis or for knowledge discovery (see Figure 6.7).

An example of the use of annotations in software outside of the visualization domain is the *macOS* file manager, the *Finder* [7]. Any file can be tagged with color or text to express a certain use or purpose and it helps to sort or filter files with a certain tag combination, making them easily accessible. Although hand-written annotation is a common practice on visualizations presented on paper [249], in the digital visualizations annotation functionality is very rarely provided. An example is the scatterplot matrix dice by Elmqvist et al. [156], where users can choose colors to encode selections of point subsets ("query layers") and follow them between animated data transitions. Others allow color customization of visual encodings [525]. These examples can be seen as a preliminary form of annotation by color, but remain limited.



Figure 1. Our dashboard prototype with context aware annotations. (a) The main dashboard contains icons with the number of annotations for each data-point (b). Bellow the list of all dashboard annotations in (c), is a set of recommended annotations (d) emphasizing their similarity.

Figure 6.7: Figure of Business Intelligence (BI) dashboard by Elias & Bezerianos [153]

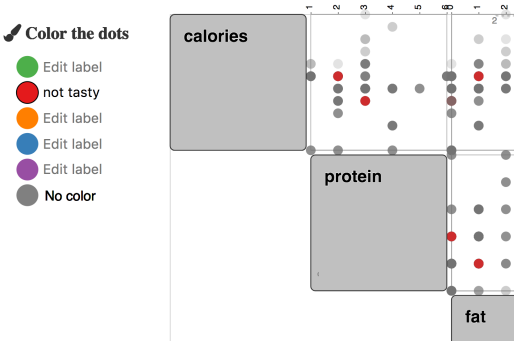


Figure 6.8: Annotation feature in DcPairs

Color hues, as discussed before, when assigned by default to certain attributes or alternatives, may prime the user. Instead, a user-driven use of colors can let the user to enrich data with personal semantics. To this end, DcPairs's annotation feature utilizes a color palette (see legend in Figure 6.8) where the user can also add textual descriptions corresponding to each color [1]. By clicking on one of the colored circles in the palette (labeled as "Color the dots") the user can tag as many alternatives as needed. This way DcPairs helps the user to document insights about the dots (e.g., to mark preferred alternatives) and add external information about the data. For example, in Figure 6.8, in a dataset of cereals, the user can annotate as "not tasty" the ones that she has tried in the past and did not like their taste. Beyond the use of color, many other possible annotation features are possible, e.g., the use of textual labels (that are likely to produce visual clutter in the DcPairs design). Tagging items with color is not a new feature in user interfaces. However, in most visualization systems the color variable is usually used as default encoding and it is rarely left to the user.

6.1.5 Communication of uncertainty

As shown in Chapter 2.4, several decision support visualization systems allow users to combine multiple attributes into a single aggregate score [85, 201, 364]. The users resize the attribute encodings

[1] DcPairs color palette is using the qualitative "Set1" color scheme with 5 colors defined in ColorBrewer [3]

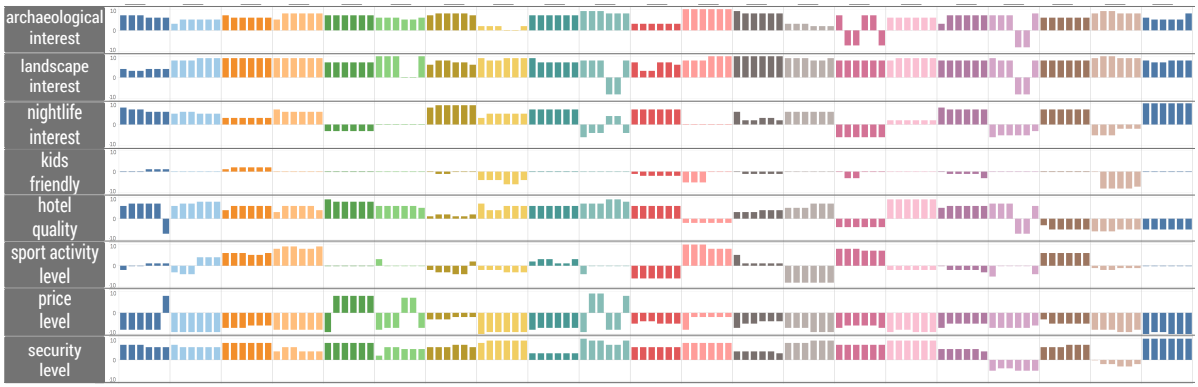


Figure 6.9: Experiment in section 3.1: Self-reported preferences for the attributes of holiday packages. The 21 participants are encoded with distinct colors. Each row shows an attribute (i.e. archaeological interest, landscape interest, nightlife interest, kids friendly, hotel quality, sports activity, and security level). Six observations (bars) per attribute for each participant (timeline order). Positive-negative values indicate direction of their preference, i.e., whether they prefer the attribute to be high or low.

to indicate attribute importance, and the system computes the aggregated score based on their expressed preferences and suggests the optimal alternatives. This approach is an interactive implementation of the "weighted additive (WADD)" method described in **??**.

Nevertheless, computations based on users' self-reported preferences can be particularly noisy. In the experiment of chapter 3, participants did not always choose the optimal alternative according to the preferences they declared. Also, the preference-based metrics showed large variability in responses, and they were not sensitive enough to capture differences across techniques. It is possible that participants may not be able to perfectly express (or be aware of) their criteria preferences. As also suggested by Pajer et al. [364], people often have a vague intuition about the relations among the weights rather than precise values. Figure 6.9 shows the raw data of the self-reported preferences in the holiday package experiment in chapter 3. Each of the 21 participants is shown with a distinct color , and each row shows an attribute. Participants were asked to rate the importance of an attribute 6 times (before and after the 3 choice tasks). Each cell shows the 6 successive ratings as a mini bar chart. If participants had had stable preferences, each cell should show bars of the same height and direction. But, as we can see, most participants appear to change their ratings over time.

We are faced with two main types of uncertainty when it comes to user preferences: the uncertainty of the user about her own preferences; and the uncertainty of the system about both the user's preferences as well as the aggregation of the user's preferences into a single score.

To better communicate user's uncertainty, DcPairs utilizes a continuous gray scale (known as a less precise color scale [495]) to encode the weight of importance of each attribute (Figure 6.10). This visual encoding is not intended to show precise values but rather to express relative



Figure 6.10: Attribute encodings in DcPairs

differences among the attributes. For each attribute, the user specifies the importance with a slider (Figure 6.10). As suggested by Matejka et al. [320], a continuous slider is used (without any labels or tick-marks) to not bias the reported preferences. Depending on the requirements of the decision maker, higher values of an attribute (e.g., safety) can be "better" or the opposite (e.g., price). A checkbox indicates the direction as "high" or "low", in contrast to systems like Value Charts where there is no natural way of specifying negative weights.

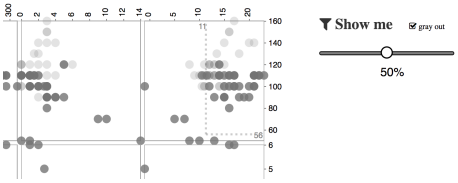


Figure 6.11: Threshold slider

To better communicate the system's uncertainty when suggesting optimal alternatives, DcPairs suggest an X percentage of optimal alternatives without indicating their exact order, as opposed to the common practice of an ordered list. A threshold slider labeled with "Show me" (Figure 6.11), (initially set to 5% best alternatives) can be dragged to determine how many alternatives the user wants to examine. The check-box next to the slider in Figure 6.11 indicates the two modes of the filtering result. The "gray out" mode filters by graying-out dots (shown in Figure 6.11) whereas he "remove" mode filters by hiding the dots. In the "gray out" mode the user can still see and retrieve information about the filtered alternatives, whereas in the mode "remove" the alternatives disappear (a feature that can help reduce clutter for large datasets with more than 100 alternatives).

6.1.6 Attribute scaling

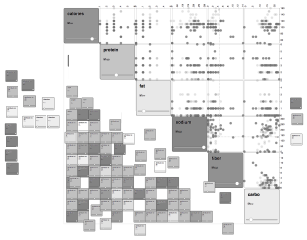


Figure 6.12: Attribute Map overview. See Figure 6.13 for a larger image.

Decision making often involves a high number of attributes to consider. For example, when purchasing a property, one has to consider price, size, the number of bedrooms, number of bathrooms, location (distance to work, schools, public transport, neighborhood), etc. Most visualization decision-support systems do not scale well to many attributes, mostly due to the color limitations explained before. The traditional scatterplot matrix also has limitations when scaling up to a dozen attributes in a large screen [347]. Nevertheless, even if a high number of attributes can be considered, during decision making not all attributes are equally interesting. For example, if a user is searching for a red car, she will most likely not be interested in the distribution of car color in comparison to consumption.

DcPairs extends the traditional pairs plot to manage hundreds of attributes, by adding a *attribute map* overview feature (bottom left in Figure 6.12). The DcPairs attribute map contains all attributes that do not appear in the diagonal of the pairs plot, and the user can drag and drop them on the diagonal to have them included in the pairs plot. Moreover, it is possible to arrange the attributes in the overview in order to sort and cluster them manually as needed. The user does not necessarily need to interact with all attributes, but the overview offers a quick glance of the number of attributes and their relative importance (gray scale color map). Clicking on the attribute magnifies it for more precise

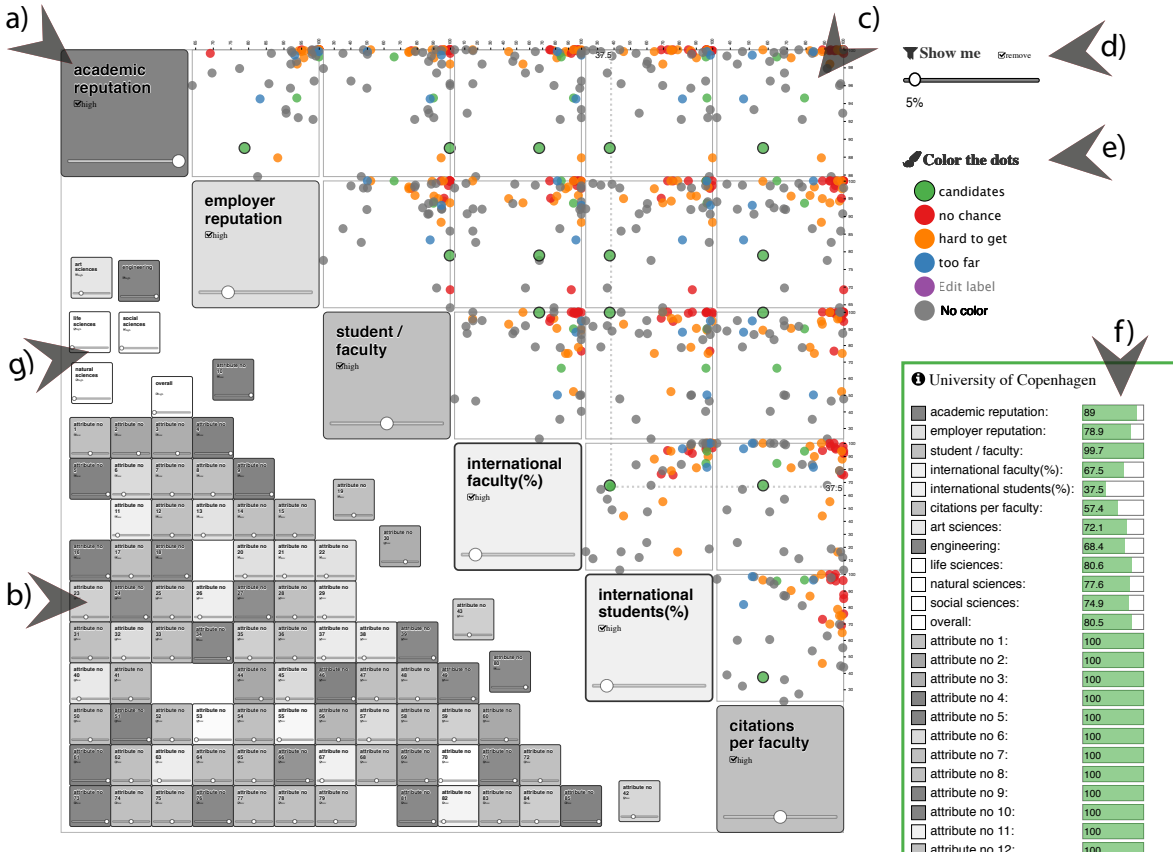


Figure 6.13: Use Case Scenario of DcPairs using QS university world rankings 2013 dataset

manipulations (e.g., changing the direction check-box).

6.1.7 DcPairs use case

A use case scenario is presented to demonstrate how annotations in DcPairs could be used to support the decision process. The scenario uses the "University Rankings" dataset from LineUp's webpage [10], containing the "QS world rankings 2013" of 906 institutions. The official ranking is based on the weighted sum of the following attributes: "academic reputation" (40%), "faculty student ratio" (20%), "citations" (20%), "employer reputation" (10%), "international faculty" (5%), and "international students" (5%), which are represented as boxes in the diagonal of the matrix in Figure 6.13 **(a)**. The dataset contains additional attributes expressing the reputation of the university in "arts", "humanities", "engineering", and "life and natural sciences". Those additional attributes are not considered in the default ranking but appear in the attribute map (Figure 6.13 **(b)**). The extra rectangles in Figure 6.13 **(b)** demonstrate the system's scalability and are not part of the real dataset.

In a scenario similar to the one described in the LineUp paper [201], Vangelis, a prospective undergraduate student from Athens, is searching for universities to apply for. He loads the "University

Rankings" dataset into DcPairs and takes a look at the gray values and slider position of the attribute boxes in the diagonal (Figure 6.13 **a**) to understand how the default score is computed. At the beginning all dots are gray (contrary to what the Figure 6.13 shows). The attribute map (Figure 6.13 **b**) gives him a quick overview of the number of the attributes and their relative importance [2]. He sees that most emphasis is laid on academic reputation, about half on citations and on the faculty-student ratio, less on employer reputation, and a lot less on international faculty and students. Even though the subject areas are deemed unimportant (signified by the white color in Figure 6.13 **g**), Vangelis is mainly interested in engineering and, to a lesser extent, in art subjects, so he drags the first slider to the maximum and the second to some low value.

Having a limited budget for application fees, Vangelis can only select three universities to apply to, so his the strategy is to choose less competitive but still suitable institutions to increase his chances of acceptance. He moves the threshold slider (Figure 6.13 **d**) to filter out all but the top 1% institutions, that he assumes are the most competitive ones, and tags them with red and a newly defined label "no chance" (Figure 6.13 **e, c**) . He moves the slider to 2% and labels the yet untagged ones as orange and "hard to get". Vangelis finally sets a 5% threshold to see institutions he considers as more realistic. He notices that most attributes are not really correlated, apart from "international faculty" and "international students" that show a weak correlation. That means he needs to consider each attribute individually. He utilizes the inspector (Figure 6.13 **f**) to retrieve detailed information, and he tags interesting institutions in green color (label "candidates"). He also uses the orange and red dots as a baseline to identify candidates similar to the top institutions.

Another aspect Vangelis cares about is to find an institution at a location that will allow him to travel to his family from time to time. This information is not in the dataset, but he extracts some locations from the institution names and others by searching online. He tags the ones extremely far from his hometown (e.g., in China) with blue and the label "too far". Using the color tags he finds his three favorite "safe" choices: the " University of Copenhagen", the "University of Manchester" and the "Kings College". Nevertheless, he decides to also apply to the "University of Oxford" (in red) trying his luck with at least one of the top institutions.

Vangelis is browsing the attribute map overview to identify other attributes that could affect his decision. Since he is interested in low tuition fees, he changes the default direction-check box of the "tuition fee" attribute [3] from a "high" to a "low" direction and drags it to the diagonal detailed view to see it paired with "academic reputation". He observes that "fee" is correlated to "reputation" in most cases, but that a few top institutions do not follow this trend and have lower fees. Vangelis interprets this as a sign that, in general, the last follow a friendlier policy towards student's budget and reconsiders his choices.

[2] The usual initial view of DcPairs is with attributes of equal weights (with the same gray color as shown in Figure 7.4). However, it is also possible to give some default attribute weights, which is the case here, in order to indicate the weights considered by the world ranking standards.

[3] Unlike the previous attributes, the tuition fee is not part of original "QS world rankings 2013" dataset. It is used here to serve the scenario of the use case.

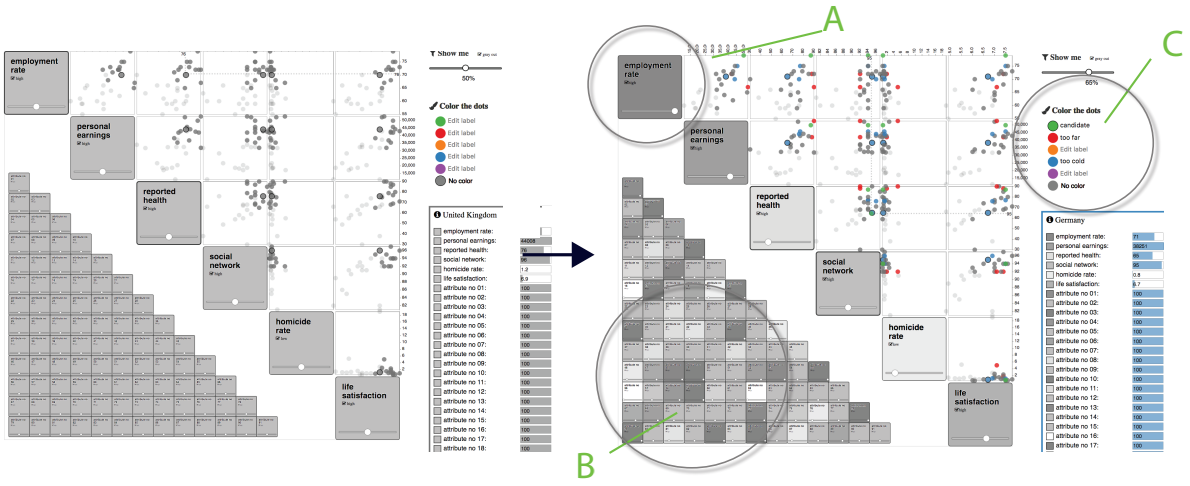


Figure 6.14: DcPairs overview with countries dataset from chapter 3. [LEFT] Initial system view before the user assigns importance weights. Identical gray boxes indicate attributes of equal importance. [RIGHT] Attributes are shown as shaded boxes; on the diagonal for a detailed view (A) others in an attribute map overview (B). Attribute weights are shown in a continuous gray scale ( ) expressing preferences as relative tendencies rather than precise values. The user annotates the countries (dots) using self-authored color-coded labels (C). Green dots show the favorite choices.

6.1.8 Summary

This section discussed a problem in the design of current decision support visualizations: the use of colors to differentiate attributes. This approach facilitates mapping of attributes across multiple coordinated views, but it has certain limitations: colors often communicate semantics (e.g., red stands for "danger") deemed to influence the user's preference, and qualitative color palettes have limited scalability. This section described a system that uses an alternative approach, named DcPairs (shown in Figure 7.4). DcPairs is a pairs plot system that offers a compact overview of the decision space, using visual encodings to communicate both uncertainty and suboptimal preference elicitation. Instead of encoding the identity of attributes, DcPairs uses colors for user-authored annotations to support the decision making process. A use-case scenario of a prospective undergraduate student choosing a university from the "QS world university ranking" dataset illustrated the functionality of the system.

DcPairs is a system currently under development. The future steps consist of (but not limited to): 1) the full implementation of a generalized pairs plot to include categorical attributes; 2) more flexible decision space For example, paired plots can be displayed anywhere in the space, by performing this interaction: when the bottom-right corner of one attribute box touches the upper left corner of another, a new plot is triggered (quantitative or qualitative based on the type of the attributes); 3) the empirical evaluation of the use of annotations in the decision making process.

The rest of the chapter will focus on another aspect of how to improve decisions with systems such as the DcPairs: their ability to facilitate unbiased decisions. The next section will review the current

debiasing techniques that are applied to help people make judgments based on rational principles.

6.2 Debiasing Techniques

Careful personnel selection or increased expertise do not seem to resolve the problem of cognitive biases [103]. Most cognitive biases appear to be quite robust even for people of high cognitive sophistication (intelligent, open-minded, or with tendency to engage in effortful cognitive activities, etc.) [506], or people with high domain expertise [395]. Instead, there are cases where domain knowledge was shown to amplify the bias. For example, professional intelligence agents who make decisions about national security issues appeared to be more vulnerable to biases than college students [395]. Cognitive biases exist in many real-world contexts such as business, medical [117, 199], legal, or military settings [352] and most strategies that have been employed to mitigate them appear so far to be rather ineffective [30, 164, 262, 411].

Simply telling people about the bias and advice them not to succumb to it is usually not effective [30]. Similarly, a strong warning message can work only under the condition that the decision maker recognizes when the bias occurs [446]. Hence, more sophisticated methods of debiasing need to be employed. This section discusses basic techniques for cognitive biases mitigation. Some debiasing techniques focus on educating and training people for specific biases, others on motivating people to perform better, and others on the re-design of the environment to avoid predictable biases.

6.2.1 Educational debiasing techniques

The most common debiasing techniques are to educate people with tutorials, extensive training sessions or simulations [199] in order to use certain approaches to solve a given puzzle. Although most attempts in mitigation training met with failure [164, 262, 411], some showed promising results [103].

"Training in rules" techniques examine the benefits of extensive training in economics (e.g., normative theory) [293], social and natural sciences [298], and statistics (e.g., law of small numbers) by combining abstract principles with concrete examples [92, 172]. "Training in rules" techniques are usually more effective for people with some rudimentary knowledge, and often weaken over time [171]. "Consider the opposite" technique explicitly instructs decision makers to consider the opposite alternative in order to widen their narrow sample of evidence and reduce biases such as overconfidence, hindsight or anchoring [30, 349]. However, this technique can backfire; when people were asked to list multiple ways in which some past events could have turned out, the difficulty to consider too many alternatives amplified some biases [411].

Computer games appear to be one of the most effective debiasing approaches. They provide immediate feedback, structured learning environments and tailored instructions based on performance [103, 143, 318, 341]. For example, participants in Clegg's et al. [103] experiment (Figure 6.15) navigated an avatar through several rooms with interactive puzzles, designed to elicit situations in which people are prone to biases (e.g., fundamental attribution error, bias blind spot and confirmation bias). Before

entering a room, participants were provided with infographics describing the bias and its possible mitigation strategy. Then, participants applied their gained knowledge into practice by solving the puzzles.

The results showed that the video game training successfully reduced the biases both right after the session and eight weeks later. The control condition provided a video with the same information, real-life examples and humor to maintain engagement, but it was outperformed by the game. According to Morewedge et al. [341], whose experiment in video games confirmed Cheng's findings, the key difference between a computer game and a video is the personalized feedback and the hands-on experience. Although narratives have also been proposed as an effective learning method in general [139], empirical results in video game studies do not yet offer support for the benefits of narratives in bias mitigation in particular [318].



Figure 6.15: Clegg et al. [103]

6.2.2 Motivational debiasing techniques

Other debiasing techniques try to increase peoples' motivation in order to put more effort in solving the task. The basic approach to increase motivation is through the use of incentives or accountability.

Arkes [30] states that the automatic nature of most cognitive biases should make people unresponsive to incentives, and that their only effect would be to motivate people to perform their "suboptimal behavior with more enthusiasm". For example, Fischhoff et al. [166] showed that asking people to wager actual money based on their confidence levels did not manage to mitigate overconfidence. The effect of financial incentives on human performance is debated among economists and psychologists. Economists generally presume that performance-based monetary rewards make people work more effectively, while psychologists typically argue that intrinsic motivations are better in producing a steady effort especially when making spontaneous rational choices or applying a Bayes' rule [82]. A review of 74 experiments showed that financial incentives can help some cognitive biases which require more effort, but this does not apply when the "know-how" and analytical skills are an important factor [82]. Interestingly, in choices under uncertainty incentives do not affect mean performance, but can reduce variance in responses [82]. Moreover, in biases that are sensitive to a favorable 'self-presentation', incentives give more realistic responses (e.g., be less generous to others or less risk-taking when gambling) [82].

Making people accountable for their decisions is another way to motivate them to put more effort. The assumption behind accountability is that people expected to justify their responses to others, anticipate possible flaws in their reasoning [76]. However, this technique implies that the person needs to be able to recognize a mistake, which is not always possible. Unlike monetary rewards,

accountability can itself bias responses to be more favorable to the audience of interest [76].

6.2.3 Computation-aided debiasing techniques

An alternative way to reduce cognitive biases would be to eliminate human judgment and incorporate statistical models based on past data – known as actuarial methods [129]. Actuarial methods are often very effective, for example in certain clinical judgments that systematically suffer from judgment inconsistencies [129]. Moreover, automated decision support systems (DSS) can run consistency checks (e.g., on attribute weights or probabilities) [256] or incorporate normative algorithms into the decision making process which would be otherwise too difficult, if not impossible, for human beings to compute [124]. However, automation can also trigger new cognitive biases when experts over-rely on recommendations from automated systems despite clear indications of disconfirming evidence (e.g., pilots who accept significantly sub-optimal flight plans from autopilot systems) [124]. Computational decision making lays beyond the scope of this dissertation, but remains a promising approach when used alone or in combination with human judgment.

6.2.4 Group-based debiasing techniques

Using the "wisdom of crowds" to debias critical decisions can also be a solution that has been applied to problems such as diagnosing cancer and financial forecasting [351]. The most effective way is for each member of a group to first form a judgment independently. Then, a final decision is the aggregated judgments [232], where the individual suboptimal strategies become less influential [438]. On the other hand, when group members form their judgments during the group discussion, individual biases are not toned out and the bias in the final decision can be rather amplified [380]. However, there is some recent contradicting evidence showing that deliberation and discussion can improve collective wisdom when responding to general knowledge questions (e.g. the height of the Eiffel Tower) [351]. An alternative way to extract better answers from large groups is to ask the members to consider both the right answer and the possible most popular answer. The variation between the two aggregate responses indicates the correct answer (known as the "surprisingly popular" algorithm) [385]. Along the same lines, the use of online rating reviews to make informed decisions, e.g., to decide which digital camera to buy, has often been investigated [300] and there is evidence that financial choices of a social network, both individual and aggregated, can help older adults to overcome their biases in financial risk taking [531]. Notwithstanding, group interactions are known to trigger other cognitive biases such as shaping favorable judgments for the group one



Figure 6.16: Group decision making. source:pixabay.com

belongs to [79, 377], and group decisions are often biased towards conformity [169] or polarization [344].

6.2.5 Design-based debiasing techniques

An alternative debiasing approach is to "*debias the environment instead of the judge*" [279]. Klayman and Brown [279] conducted a study in diagnostic reasoning where participants had to distinguish fictional diseases. The study gave descriptions of the diseases in textual format and varied only the structure of the information provided without attempting to modify peoples' judgmental processes. One group of participants read about each disease separately, and another group saw the information about the two diseases juxtaposed to highlight distinctive features. Juxtaposed representation led to more accurate judgments. Another approach suggests that information presented in a disfluent format can reduce a confirmation bias [227]. For example, jurors appeared to give less confirmatory verdicts when reading a summary of a crime which required processing difficulty [227]. Medical research, though, suggests that an environment that offers more information, such as checklists, deliberate practice, and immediate and focused feedback can benefit clinicians in avoiding diagnostic biases [199]. However, despite the discernible effort in the medical field to identify and reduce prevalent diagnostic errors, there is little to no empirical data to demonstrate notable improvement in decision making performance [117, 199].

6.2.6 Summary

This section reviewed debiasing techniques that try to address the distorting effects of cognitive biases. Several approaches have been outlined, namely educational, motivational, computation-aided, group-based and design-based. The most effective techniques seem to be the educational ones, in particular when people received training session through a video game. However, their experimental design did not include testing for generalization across multidisciplinary contexts or long-term transfer of knowledge. For example, it is unclear whether an analyst who successfully solved a confirmation bias puzzle can improve her reasoning during actual visual analysis. Besides, access to extensive training is not always a feasible solution.

This chapter explores a design-based debiasing approach which focuses on transforming the environment of the decision maker. The studies described in section 6.2.5 provided evidence that restructuring the way information is presented can mitigate some cognitive biases. Nevertheless, those studies examined onl y textual representation formats, and, although visualizations have been suggested as promising interventions for improving rational reasoning [536], it seems that there is no empirical evidence on whether visualization systems can mitigate cognitive biases. The next section will focus on an empirical investigation of how a visualization system could alleviate a cognitive bias.

6.3 Deletion Experiment: Alleviation of the Attraction Effect

Although DcPairs considers many factors to improve users decisions (e.g., annotation feature, the effect of colors, preference elicitation methods), it still has an important limitation: inferior, dominated data points, e.g., less competitive universities, can affect the attractiveness of other data points (Figure 6.17). This section investigates how a visualization designer can alleviate the attraction effect (observed in chapter 5) in visualization systems that target decision support.

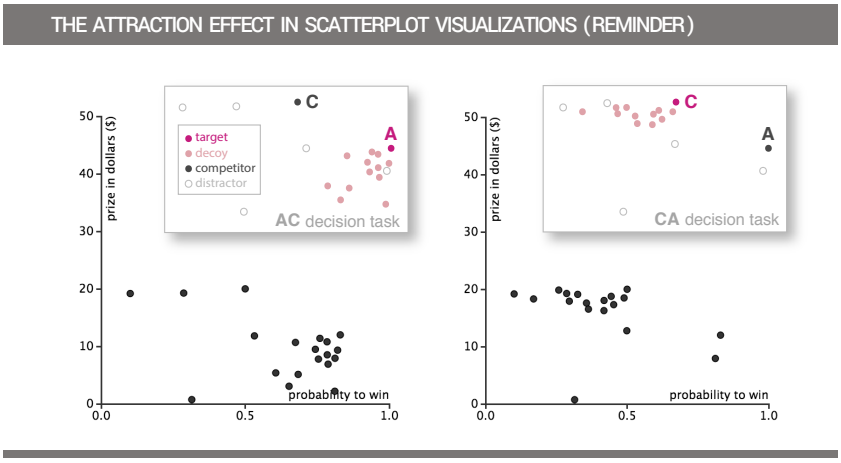


Figure 6.17: Illustration of the attraction effect in scatterplot visualizations. The position of the decoys affects the attractiveness of the **target** (A for the left plot, C for the right plot) – reminder from Bets experiment in section 5.5

6.3.1 Design rationale

The simplest way to eliminate the attraction effect would be only to show the Pareto front, i.e., to hide all dominated options (Figure 6.18 A). However, hiding the points assumes that the system has full knowledge of the user's choice criteria, which may not be the case in practice. In addition, dominated options can help understand dataset trends, and may provide useful context when making decisions. Thus, such debiasing techniques should only be available as options and activated on demand.

Another approach towards a possible alleviation of the attraction effect could be to try altering the visual appearance of the data points. A visualization designer could consider visually highlighting the Pareto front (Figure 6.18 B) or de-emphasizing dominated options (Figure 6.18 C) using color or opacity. Such techniques could have an effect and remain to be experimentally tested. Nevertheless, these approaches assume that the attraction effect is mostly a perceptual bias. It is indeed possible that the attraction effect can also have perceptual origins. Its robustness though in non-visual stimuli (oral, text, etc.), indicates that the attraction effect also has strong cognitive origins and thus pure perceptual solutions may not be enough to alleviate the bias. Besides, even if the bias can be alleviated perceptually, design methods which can be applied only to dominated points do not necessarily resolve similar context based problems. For example, there are other types of data points, not necessarily

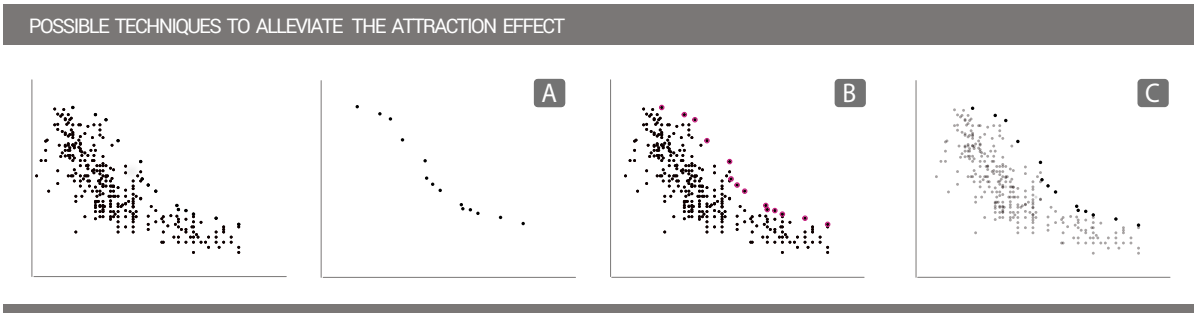


Figure 6.18: Possible techniques to alleviate the attraction effect in scatterplots: remove the dominated points (A), highlight the Pareto front (B), de-emphasize the dominated points (C).

dominated, which can still affect the attractiveness of other data points (see other context-generated biases in Appendix A such as compromise effect, phantom effect, etc.).

The approach of this chapter towards the alleviation of the attraction effect is to encourage a more effective decision strategy. Well-identified decision making strategies, such as the "weighted additive (WADD)" or the "Frequency of good and bad features (FRQ)", are known to be useful to decision makers when dealing with complex multi-attribute choice tasks that can depend on personal preferences [380], or that involve several iterations over a large number of attributes and trade-offs (decision strategies were described in section **??**). In the "Bets" experiment the choice task was simpler, consisting only of two attributes and a number of clearly inferior alternatives. Using a decision strategy for such a simple task may seem unnecessary. However, it is also possible that a more structured way of dealing with decisions can trigger more rational responses.

An interesting decision strategy that seems to have some common elements with the attraction effect's choice task is the "elimination by aspects (EBA)" proposed by Tversky in 1972 [471]. In EBA, the decision maker first rejects all the alternatives that do not satisfy her choice criteria to end up with the alternative of her choice. Similarly, a typical attraction effect choice task is divided into two subtasks: the decision maker is expected to first recognize the dominance by rejecting the decoy(s), and, second to choose between the two trade-off choices, target and competitor [116]. As explained in section 5.1.1, it is often claimed that the first step of the dominance recognition process causes the bias. There is also a small indication in favor of this argument from our inconclusive "Real experiment": when people did not perform an active dominance recognition task, the effect was not observed (section 5.4). The key challenge is how to ensure that the dominance recognition task does not affect the final decision. In the EBA strategy, the unwanted alternatives are removed from the choice set. If, in the same way, the visualization system allows the decision maker to visually delete the unwanted alternatives, could this reduce the attraction effect?

The following experiment named "Deletion" investigates whether enforcing an EBA strategy, in which participants interactively first delete unwanted alternatives, will make participants less vulnerable to the attraction effect.

6.3.2 The deletion task

Deletion is a low-level analytic task [102] that it is usually omitted from most visualization taxonomies [20, 402, 445, 504, 526]. The term **deletion** in this dissertation is defined as the task of removing one or more data cases from a visualized dataset. Two types of deletion are defined: manual deletion and filter-based deletion. A **manual deletion** consists of removing one or more data cases that have been explicitly identified by the user. A **filter-based deletion** consists of removing all data cases that match some user criteria. In other words, filtering is a type of deletion, but deletion is not necessarily a filtering task. A simple example to illustrate the difference is the following choice task in a house dataset. A decision maker who wants affordable houses, can remove data cases based on their "price" attribute (filter-based deletion). A decision maker who does not like the photo of a house (or due to other criteria not included in the dataset) can remove the specific house data case (manual deletion). The filter-based deletion creates a rule in which all data-cases that meet it are removed (e.g., all expensive houses). In a manual deletion, the data cases with similar or even identical house photos will not be removed because the system is not aware of the removal criteria. Most decision support visualizations reviewed in section 2.4 allow filter-based deletion but not manual deletion. For example, in the LineUp system, the user can not directly remove a specific row from the table, but she can remove rows by specifying attribute filters [201].

A manual deletion of data cases may not be a very common interaction during visual analysis tasks. For example, in a dataset with two data points A and B with values $X_A > X_B$ for a given attribute X, if a user wants to remove B as an outlier because of its high X value (e.g., to reduce clutter), she will also want to remove outlier A. This is a common filter-based deletion based on a value threshold. Deleting though only the outlier B without a link to other dataset features, would be a rather uncommon task or a source of confusion. An explanation of why deleting data cases manually is uncommon in visual analysis may be the way interaction is defined in the visualization field: a change or adjustment of the visual data representations which it does not usually involve users entering data into the system (as opposed to the HCI field) [526]. In a manual deletion the user performs an action based on knowledge not derived from the dataset, by signifying that "this alternative is not suitable" or "this information is irrelevant". Therefore, a manual deletion, to some extend, can be a type of user input.

Nevertheless, unlike visual analysis, manual deletion can be very useful in decision making. First, it is very likely that a decision maker will make a decision based on criteria not included in the dataset, e.g., to rule out solutions that she tried in the past and knew that they do not work. Moreover, even if the decision maker wants to make a decision based on information that exists in the dataset, it does not necessarily mean that she can determine at a given time a precise rule out of these criteria. For example, the house buyer may delete a house she finds too expensive for having only one bathroom and no other particular appealing attribute. This does not necessarily imply that she wants all expensive houses with one bathroom to be removed (e.g., she could compromise with a house with other room surplus and appealing big garden). Choice preferences can evolve during exploration and are often formed progressively based on the availability of the alternatives. Manual deletion can also be important in

cases where a decision maker wants to apply certain decision strategies, e.g., the "elimination by aspects (EBA)" discussed before. In the following experiment named "Deletion", another possible benefit of deletion will be investigated: its potential in the alleviation of decision biases. The assumption is that by explicitly "cleaning up" the decision space from irrelevant information can help a decision maker to make a more rational decision.

6.3.3 Experiment task

The Deletion experiment replicates the "Bets" experiment conducted in section 5.5, giving participants twenty choice tasks of lottery tickets each defined by the *probability* of winning, and the *prize* (the amount that can be won). Each choice task consisted, again, of a target, a competitor, decoys and distractors as described in sections 5.5.1.2 and 5.5.1.3.

A choice task, as defined in section 2.1.2.2, is the task of *finding the best alternative among a finite set of alternatives*, but it does not specify the procedure that a person follows in order to complete the task. The current experiment examines two different procedures to conduct a choice task: the procedure followed in "Bets" where participants directly selected the alternative of their choice (shown in Figure 6.19), and the procedure of first deleting all the unwanted alternatives to end up with the alternative of their choice (shown in Figure 6.20).

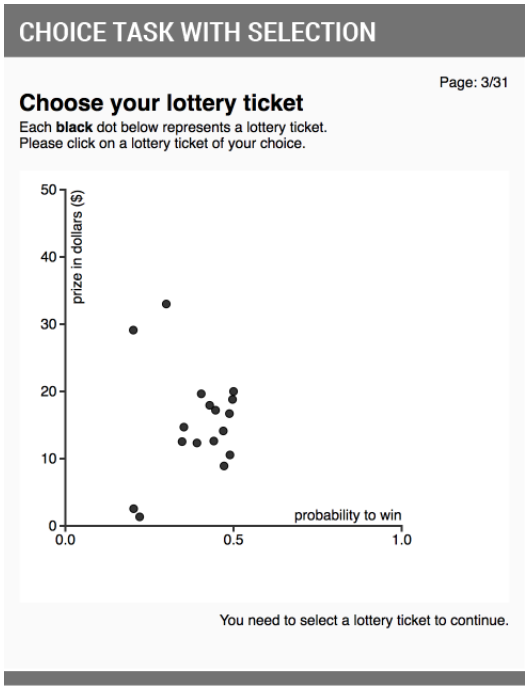


Figure 6.19: Deletion experiment: the task in selection condition (baseline)

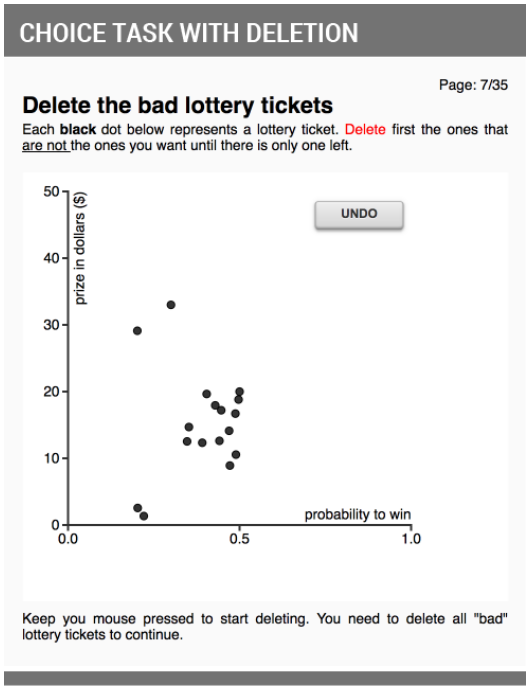


Figure 6.20: Deletion experiment: the task in deletion condition

6.3.4 Interactions

In the selection condition, the interaction was the same as in the "Bets" experiment. Dots were highlighted when hovered. Hovering a dot also displayed horizontal and vertical projection lines, and the dot's X and Y values were overlaid on the axes. After a dot was clicked, its color changed and the participant was asked to confirm her choice by clicking on a button at the bottom of the page. To undo a selection, the user could click again anywhere else apart from the confirmation button.

In the deletion condition, the hovering feature was the same as in the selection. While the mouse was pressed, a red outline was displayed around the scatterplot to indicate that the user is on deletion mode. During deletion mode dragging over a dot would remove it. The user could also delete dots by instantly clicking on them. The deletion area was substantially smaller than the size of the dot. This made the deletion more tedious, but it was necessary to prevent errors since some decoys were very close to the target and to each other. To undo a deletion, an "undo" button was offered to recover the deleted dots.

Choosing an item by deleting unwanted items is not a common interaction in user interfaces. So in order to prevent accidental deletions, participants were given the following piece of instructions "*During the tutorial* [4]*, you clicked on a dot to select it. You will not do this anymore. First, you have to delete the dots you do not want. The last dot you will leave will be the one you select.*" (page 1) and then a *Deletion Training and Pre-test* which is shown in Figures 6.21, and 6.22. Participants had to be able to delete given red dots and leave the black dots untouched in order to proceed to the experiment. The dot patterns presented in Figures 6.21 and 6.22 were selected carefully so as not prime users by showing an attraction effect pattern and to also include a precise deletion of a red dot being very close to a black one. If the participant deleted by mistake a red dot, she was instructed to press the undo button to redo the task from scratch.

6.3.5 Experiment design

The Deletion experiment followed a mixed design. The between-subjects factor was the *interaction technique* (selection or deletion). As in the "Bets" experiment, the within-subjects factors were the *task pair* (10 pairs of matched tasks), and *decoy position* (on probability or prize).

6.3.6 Participants

The population sample consisted of 203 crowdsourced contributors (101 for the deletion condition and 102 for the baseline selection [5]) who submitted valid responses. A summary of participants' self-reported demographics is shown in Figure 6.23 (map and bar charts). As it can be seen, participants tended to be educated young male adults.

[4] The tutorial refers to the test on probabilities and scatterplots given in both conditions (see procedure section).

[5] The planned size was 200 which is hard to precisely control in the on-line platform.

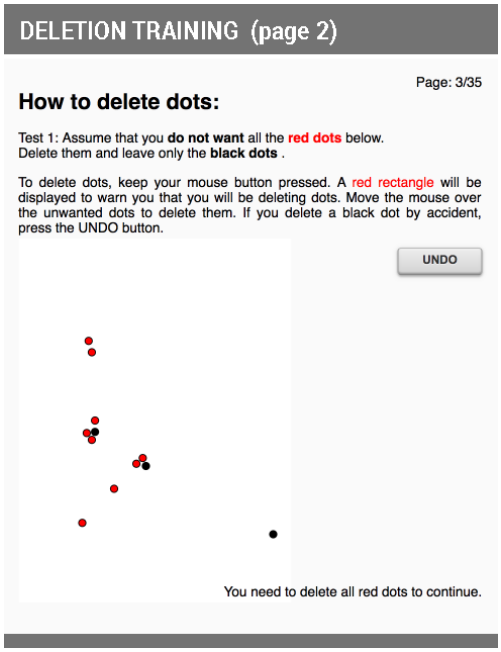


Figure 6.21: Deletion training (part2)

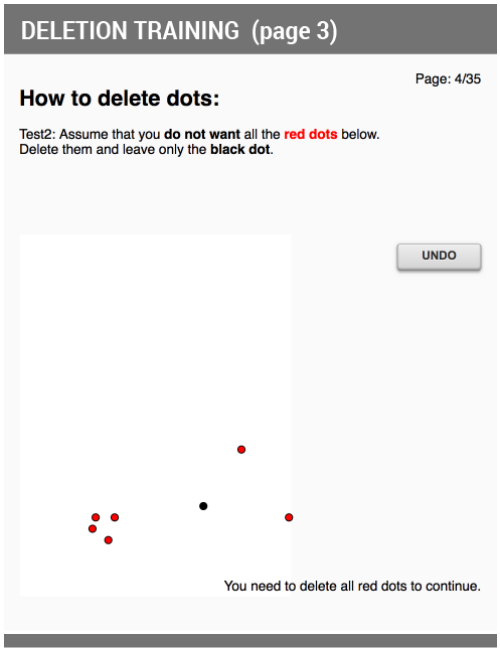


Figure 6.22: Deletion training (part3)

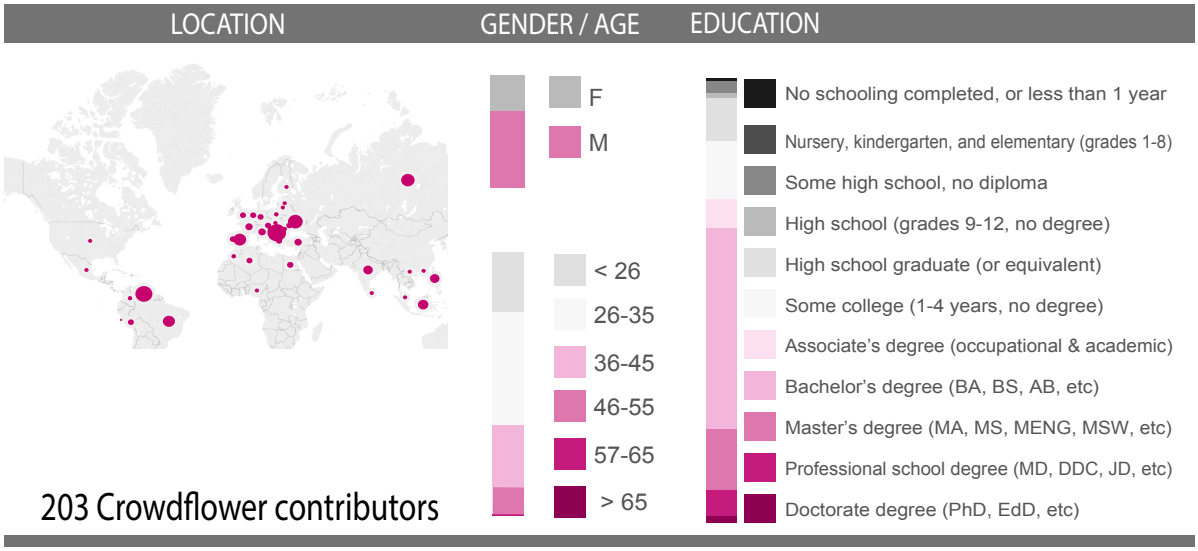


Figure 6.23: Deletion experiment: self-reported participant demographics

6.3.7 Procedure

Each participant of the baseline *selection condition* followed an identical procedure with the ones in "Bets" experiment, mainly consisting of: tutorial and a test on probabilities and scatterplots (*Training and Pre-test*), a motivation message that the bonus payment depends on the choice (*Pre-test: Motivation*), 20 choice tasks of lottery tickets presented in a scatterplot, the selection of the lottery ticket of their choice by clicking with their mouse on the corresponding dot (*Main Study Task*), and a post-test questionnaire (for more details see section 5.5.2.1).

Each participant of the *deletion condition* followed the same procedure as in the selection, with two differences. First, after the *Pre-test: Motivation*, they went through the *Deletion Training and Pre-test* (Section 6.3.4). Second, the interaction in the *Main Study Task* was a deletion instead of a selection (Section 6.3.3).

The duration from the *Main Study Task* until the end lasted on average 6'28" for the selection condition and 14'07" for the deletion condition.

6.3.8 Hypotheses

The research hypothesis was:

Hr1 Following the "elimination by aspects" decision strategy while interacting with a scatterplot reduces the attraction effect.

This translates into the following statistical hypothesis:

H1 The mean attraction score of the deletion condition will be smaller than the mean attraction score of the selection condition.

6.3.9 Planned analysis results

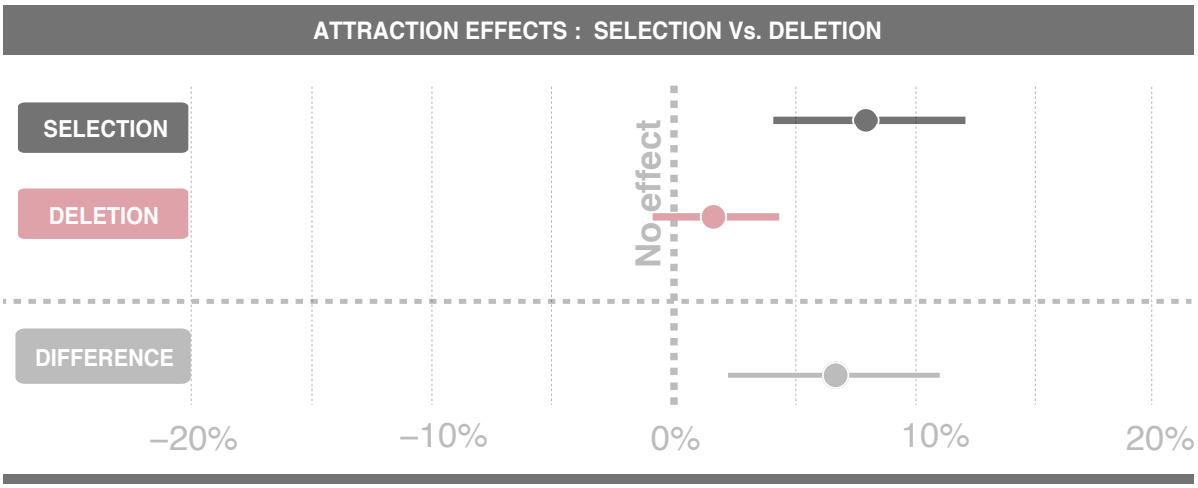


Figure 6.24: Point estimates and 95% confidence intervals for the attraction effects in Selection and Deletion conditions.

The description of the attraction score is explained in detail in Section 5.5.3. Briefly, the attraction score for each participant is based on the difference in choices made when the decoys were on prize and the choices made when the decoys were on probability. A participant whose preference for high probability or prize is independent from the position of the decoys should have attraction score close to zero.

For the baseline condition (selection), where participants directly clicked on the ticket of their choice, the mean attraction score S_{baseline} was 8%, with a 95% bootstrap confidence interval of [4%, 12%] (see Figure 6.24). Like in the previous Bets experiment, there is still solid evidence that participants were subject to the attraction effect.

For the deletion condition, where participants deleted first the unwanted tickets, the mean attraction score S_{deletion} was 1.3%, with a 95% bootstrap confidence interval of [-1.3%, 4%] (see Figure 6.24). The range of plausible mean values of the effect indicates that either participants were not subject to an attraction effect or that the effect is relatively small.

The difference $D = S_{\text{baseline}} - S_{\text{deletion}}$ in means between the two conditions was 7%, with a 95% bootstrap confidence interval of [2%, 11%]. Thus, there is strong evidence that the hypothesis **H1** is confirmed.

6.3.10 Additional results

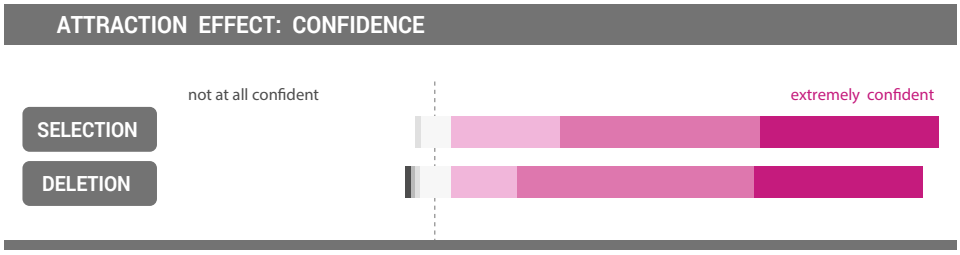


Figure 6.25: Deletion experiment: Self-reported confidence

As shown in Figure 6.25, participants were confident in their choice overall, although perhaps slightly less in the deletion than in the selection condition.

6.3.11 Discussion

The Deletion experiment confirms that following the "elimination by aspects (EBA)" decision strategy while interacting with a scatterplot reduces the attraction effect. As discussed in section 6.2, most debiasing techniques usually require extensive training (formal courses, tutorials or video game simulations [103], computation aid, etc.) with questionable effectiveness in long-term use and multidisciplinary contexts. The design-based techniques which focus on improving the environment instead of educating the user, despite the discernible effort in the medical field towards this direction, had showed limited empirical evidence of their effectiveness. The Deletion experiment can be an encouraging first step

towards design-based solutions. It showed that it is possible to provide interactions that simulate well-identified decision strategies to help users making more rational decisions.

Nevertheless, the reasons why EBA led to more rational decisions (in the sense of being consistent with axioms of rational choice theory, see Section 2.2) remain yet unclear. One important factor to consider in a future experiment is the effect of time spend on the task. In the deletion condition, participants deleting manually all the dots spent twice as much time as those in the selection condition.

The main reasons why deletion was more time consuming are twofold. First, it is the inevitable cost of enforcing a structured decision making strategy. Decision makers when using these strategies need to explicitly iterate over almost all alternatives. For instance, if the system had enforced "weighted additive (WADD)" strategy instead, the user would have to weight all attributes by importance and compute a weighted sum per alternative. On the other hand, the selection condition did not enforce any strategy. Participants could have followed in their mind a specific strategy (or a combination of them), e.g., EBA by rejecting decoys first and then deciding between target and competitor, "satisficing (SAT)" by scanning alternatives and click on the first that satisfies their needs, "lexicographic (LEX)" by deciding which attribute is most important (probability or prize) and click on the alternative that is best on that, or even they could have chosen to click at random. But in any of these possible (un)conscious processes there was no interaction requirement. Second, the deletion condition could have taken extra time due to the way the interaction was implemented. As explained in the section 6.3.4, the deletion interaction did not allow for a deletion area bigger than the size of the dots, in order to avoid accidental deletions when the decoys are too close to the target. A more flexible deletion method (e.g., with larger deletion area) could had improved the overall task time, but it would still be impossible to compete with a direct click on a single dot. Besides, a more flexible deletion could have led to more errors burdening further the completion time by the cost of undo operations.

Regardless of the reasons why deletion was more time consuming, the fact that participants spent more time in each diagram may had en effect in their reasoning effort. It is possible that this extra time allowed them to consider their choice more thoroughly and decide based on their true preferences. Conversely, it is also possible that the repetitive deletions were performed in a mechanical manner without any conscious effort. One question to consider is whether it would be possible to make the selection condition equally time consuming, for example, by freezing the screen for a few minutes. However, altering the expected way that selection usually works does not guarantee extra cognitive effort by the user. It could instead introduce further noise, for example, the delay may frustrate the user that could switch to another task. Generally, exploring ways to encourage more conscious effort from the user is an open question that is worth further investigation.

Perhaps a more plausible explanation as to why the deletion led to more rational decisions is that "cleaning up" the decision space from irrelevant information helped participants to choose up front. In addition, the process of deleting decoys may have helped them to consciously strengthen the idea that the deleted piece of information is irrelevant or unimportant.

A critical aspect of the Deletion experiment that needs clarification is that enforcing a particular

interaction or, even worst, prohibiting a direct selection, is by no means recommended as a practice. Forcing users to perform an otherwise easy task (with only two obvious good alternatives) with a tedious interaction would not make sense in a real system. The purpose of the enforced deletion was to study its effect and to encourage visualization systems to enrich the interactions offered to the user. Decision strategies like EBA would be practically useful in more complex tasks with multiple alternatives and attributes. In such tasks, the time and effort requirements would be demanding regardless of the strategy. A similar practice of using interactions to support a strategy is illustrated by systems such as Value Charts [51] which encourage the "weighted additive (WADD)" strategy by allowing weight manipulations of importance and aggregation of a score through the stacked bars. However, the users can perfectly ignore this feature and review the information using only the tabular layout.

6.4 Conclusion

This chapter examined the improvement of decision support visualizations from various angles. First, it discussed how the choice of visual variables, particularly color, can influence decisions and introduced DcPairs, a novel visualization system which supports multi-attribute choice tasks with user-authored annotations. DcPairs utilizes vague visual encodings to communicate two types of uncertainty: uncertainty of the user regarding her preferences, and uncertainty of the system regarding the suggestion of optimal alternatives. DcPairs further offers a compact decision space that allows the user to manage more than a hundred decision attributes. A use-case scenario of a prospective student who chooses a university illustrates how color annotations can assist a user during the decision making process. The potential benefits of DcPairs remain to be empirically validated.

The chapter further considered additional ways of improving decision-support visualizations like DcPairs, by reviewing methods that target alleviation of cognitive biases . The debiasing techniques were grouped under the categories educational, motivational, computation-aided, group-based and design-based. Finally, the chapter followed a design-based debiasing approach and introduced a new method for bias alleviation: the use of the "elimination by aspects (EBA)" decision strategy to alleviate the attraction effect in scatterplot-based systems. The EBA strategy was implemented using a manual deletion analytic task. In particular, the chapter empirically compared a direct selection of data points with a manual deletion of unwanted data points to end up with a single chosen one, and found strong evidence that the deletion leads to more rational decisions. This experiment is however preliminary and more studies are needed to better understand what causes this effect.

Notably, current visualizations systems that target decision support do not usually allow the users to annotate the data or manually remove unwanted information based on external knowledge. Both the DcPairs use case and the deletion experiment showcase a need to enrich interaction in visualizations to support a more liberating decision process.

Chapter 7

General Conclusion & Perspectives

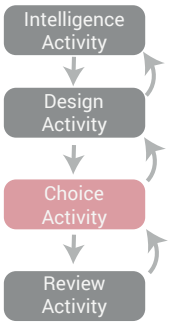
The background section of this dissertation started by observing that although visualization scholarly books acknowledge decision making as a core challenge of the visualization domain, they do not report methodologies or tasks to assess if a visualization system can assist users in making better decisions. This suggests that visualization research may not fully address the challenges of visualization-supported decision making. This assumption was later investigated and confirmed by an extensive review of decision theory and visualization literature in Chapter 2.

The most important limitation of visualization-supported decision making that was identified in Chapter 2 is the lack of *empirical* works. The following Chapters 3, 4, and 5 attempted to bridge this gap by empirically investigating decision tasks, instructions and human limitations when using visualizations, respectively. Based on the empirical findings from these chapters, Chapter 6 focused on how to improve visualization systems in a way that can help people make better decisions. The goal of the current chapter is to summarize the findings and contributions from this dissertation, draw general conclusion, and highlight potential opportunities for future research.

Specifically, **Section 7.1** recalls the previous chapters of this dissertation and summarizes their main findings. **Section 7.2** discusses important limitations in the studies summarized in the Section 7.1. **Section 7.3** outlines the individual contributions in each chapter of this dissertation. **Section 7.4** attempts to synthesize a general conclusion with respect to the thesis statement as presented in the introduction (Section 1.1). Finally, **Section 7.5** discusses some topics that can be interesting to address in future research.

7.1 Summary of Lessons Learned

Chapter 2 reviewed background work in decision theory and information visualization.



Section 2.1.1 reviewed visualization tasks: high-level (e.g., confirmatory analysis) and low-level (e.g, clustering) and identified four properties that describe their nature, and their relation with other analytic tasks. The section concluded that decision tasks are neither part of visualization task taxonomies nor formally defined. Similarly, **Section 2.1.2** presented a high-level decision task involving four stages (Figure on the right) and then defined a low-level decision task named *(multi-attribute) choice task* [1]. The *choice task* shares the same four properties as the low-level analytic tasks, but differs from low-level analytic tasks in that it serves different user goals. The goal here is not to compare values, sort, determine ranges or correlations; the goal is to select the single best among several alternatives.

Section 2.2 covered the two branches of decision theory: normative decision theory, which conceptualizes how people can make optimal choices given a set of constraints and values; and descriptive decision theory, which attempts to analyze how people actually make decisions. Descriptive theories suggest that decisions can be subjective, context-based, and liable to unpredictable variations and it provides several choice strategies (e.g.,"weighted additive (WADD)", "elimination by aspects (EBA)") that can help people to maintain consistency depending on the problem type.

Section 2.3 presented systematic errors that humans make during decision process, known as cognitive biases. The section reviewed known biases by classifying them under a new taxonomy named "FAULTY". FAULTY is organized by task to help visualization designers look at which biases may exist in their system, assuming they know the tasks users will perform (a choice task, an estimation under uncertainty, etc). The section further reviewed visualization research on cognitive biases and showed that, although cognitive biases are often emphasized as important, there is little to no empirical work that validates the existence of cognitive biases in visualizations.

Section 2.4 reviewed 38 visualization systems that target decision-support, presenting their design and the methods used to evaluate their effectiveness in supporting multi-attribute choice tasks. The review concluded that the effectiveness of visualizations for decision-support is not explicitly addressed. Specifically, there remain important limitations in their evaluation methodologies, namely: only 7 systems conducted *controlled experiments*; only 8 used *decision making tasks*; there was a lack of *sensible baselines*, and, most importantly, a lack of metrics for *decision quality*.

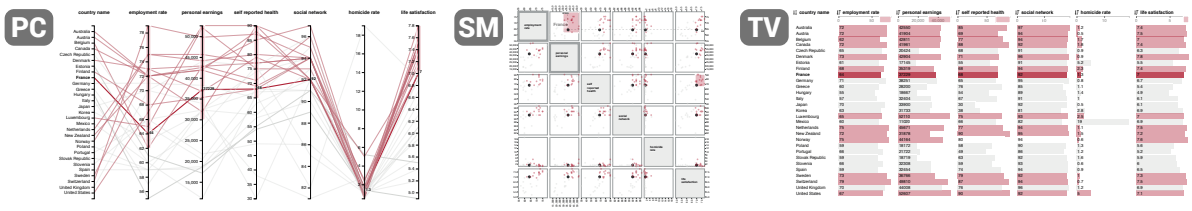


Figure 7.1: Parallel Coordinates (PC), Scatterplot Matrix (SM) and Tabular Visualization (TV).

[1] Terminology reminder: The *multi-attribute choice tasks* are called, simply, *choice tasks* when the number of attributes is relatively small (one or two) or when there is no need to emphasize the number of the attributes. The term *decision task* refers to all types of decisions tasks, including the *multi-attribute choice tasks*.

Chapter 3 attempted to address limitations in the evaluation of decision support visualizations as identified in Section 2.4. First, a *systematic analysis* articulated the link between multi-attribute choice tasks and multidimensional visualizations, investigating the extent to which multidimensional visualizations are appropriate for such tasks. The analysis identified the family of "lossless" geometric visualizations as likely the most compatible. Nevertheless, the analysis showed that we know little about how elementary multidimensional visualizations compare in terms of analytic tasks, and even less so in terms of how they support decision tasks. The chapter compared parallel coordinates (PC), scatterplot matrices (SM), and tabular visualizations (TV) (Figure 7.1) for their ability to support a decision. In this evaluation, the chapter introduced new decision metrics and outlined a methodology on how to evaluate visualizations for decision support involving both analytic and decision tasks.

Regarding the analytic tasks, SM was by far the fastest in correlation tasks and although PC is considered effective at conveying correlations, it was outperformed by TV both on time and accuracy. PC and TV were faster for value retrieval and range tasks. Regarding the decision tasks, despite the low attention it has received in the literature, TV showed a small speed advantage over the other techniques. This result was not anticipated by the analytic tasks. Otherwise, all techniques appeared to be comparable across most metrics. Moreover, participants reported a preference for TV over PC, and appeared more attached to the choices they made with SM. Regarding the metrics used, the decision accuracy metric showed a much larger variability in responses compared to the analytic results. Moreover, the indirect metric (attachment) for assessing choice confidence appeared to be more accurate than simple confidence ratings.

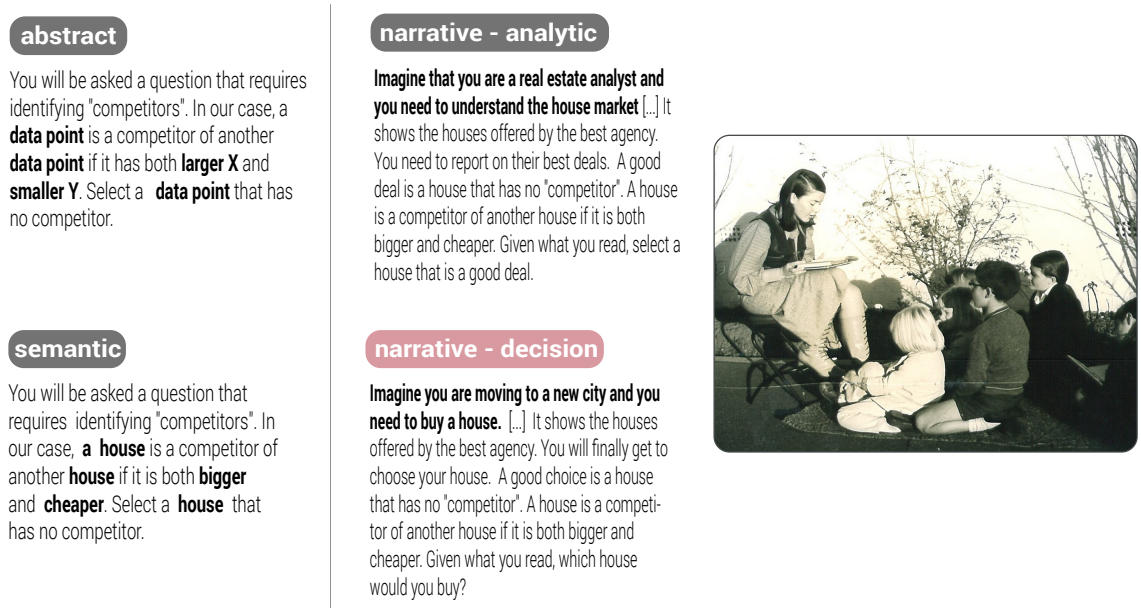


Figure 7.2: A sample of instructions given to participants in narratives experiment (Chapter 4).

Chapter 4 observed that narratives seem to be an essential and inevitable component when providing instructions for decision tasks. Therefore, the chapter investigated the role of narratives in decision tasks and possible differences between decision and analytic tasks. The chapter examined whether a narrative component can engage and motivate users to give more accurate responses or if instead longer instructions induce more errors. Crowdworkers were given i) abstract information visualization tasks without any context, ii) tasks with added semantics to the dataset, and iii) tasks with two types of backstory narratives: an analytic narrative and a decision narrative (Figure 7.2).

Contrary to the stated expectations, there was no evidence that adding data semantics increases accuracy, and further, narratives were shown to decrease accuracy. Adding semantics can, however, increase attention and provide subjective benefits in terms of confidence, perceived easiness, task enjoyability and perceived usefulness of the visualization. Nevertheless, narratives did not appear to provide additional subjective benefits. In addition, although narratives can generally be a reason for less accurate responses, it seems that is not the main reason why decision tasks are so prone to errors. In fact, narratives were also tested with the analytic framing, but participants were more accurate than when the visualization task was framed as a decision problem. The dominance-based metrics used to evaluate decision quality were not enough to help us understand this difference.

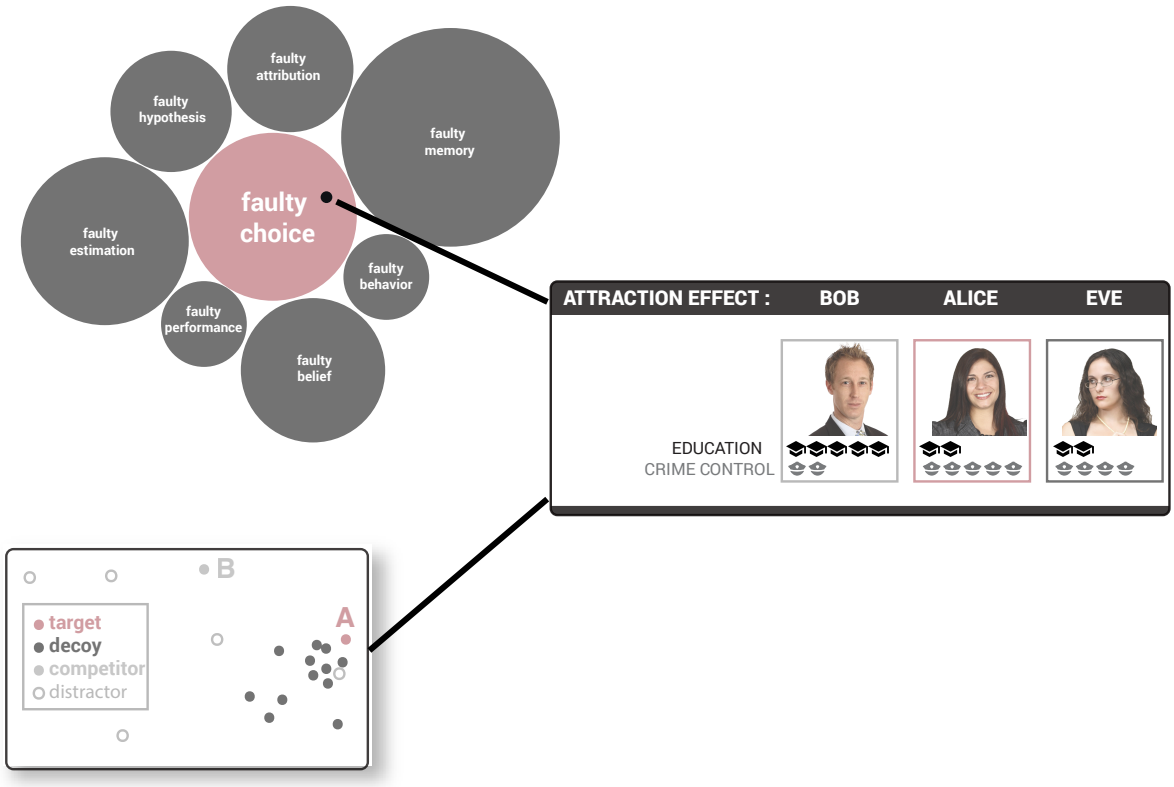


Figure 7.3: The taxonomy "FAULTY' of 139 cognitive biases and the attraction effect (the black dot).

Chapter 5 attempted to address the lack of empirical visualization works in cognitive biases as identified in Section 2.3. The *attraction effect* is a cognitive bias – one of the 23 from the *faulty choice* category in FAULTY (Figure 7.3) – that has been observed under a variety of experimental conditions, the majority of which present choice tasks as numerical tables. A few studies had observed the effect in photos, verbal instructions, or physical objects. However, there is a debate in consumer research on whether the effect indeed generalizes to non-numerical (visual) representations. Moreover, psychologists suggested that the attraction effect should disappear if more than three alternatives are added, since pairwise comparisons and dominance recognition becomes hard if numerical tables are used. By conducting series of experiments ("Gyms", "Real", and "Bets"), this chapter investigated whether the attraction effect can affect people's choices while using scatterplot visualizations.

The experiment "Gyms" suggested that the attraction effect generalizes to data visualizations. The experiment "Bets" suggested that the attraction effect can also be observed in choice tasks involving more alternatives, likely because scatterplots support fast comparisons and dominance recognition. Overall, the attraction effect study indicates that when people visualize choice alternatives, the number and position of irrelevant data points may influence their choice. This shift in preferences violates basic axioms of rational choice theory as presented in Section 2.2. This chapter used metrics of decision quality which involve more sophisticated factors that can make a choice irrational. Such metrics appear to be more informative than the preference-based metrics of Chapter 3 and the dominance-based metrics of Chapter 4. It became evident that, even in visualization, a decision can be "correct", yet irrational, in the sense that users choices can be influenced by irrelevant information.

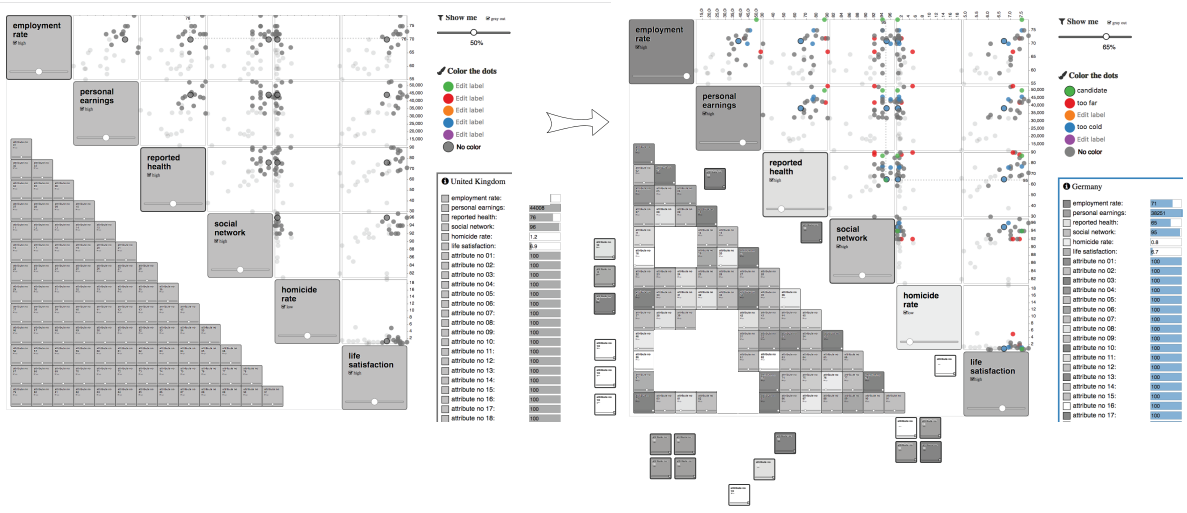


Figure 7.4: Reminder of DcPairs system described in Chapter 6. [LEFT] Initial system view before the user assigns importance weights. Identical gray boxes indicate attributes of equal importance. [RIGHT] Attributes are shown as shaded boxes; on the diagonal for a detailed view others in an attribute map overview. Attribute weights are shown in a continuous gray scale ( ) expressing uncertainty in user preferences. The user annotates using self-authored color-coded labels

Chapter 6 investigated how to improve visualizations by helping people make better decisions. The chapter identified two additional limitations of the decision-support visualizations reviewed in Section 2.4: the tools provide *limited interactions* (no manual deletion, annotation, etc.) to aid decision strategies, and they *misuse colors* in a way that limits their scalability and can impede decisions. The chapter also considered limitations that were identified in studies of previous chapters: the *uncertainty* involved in user preferences and the *cognitive biases* that can affect visualization designs. Considering all those limitations, the Chapter 6 looked for solutions in the decision strategies reviewed in Section 2.2 to support consistency. Moreover, it reviewed and classified existing debiasing methods into educational, motivational, computation-aided, group-based and design-based.

Based on this analysis, Chapter 6 examined two solutions. First, it presented a novel decision-support tool, named DcPairs. DcPairs supports user-authored annotations, uncertainty communication, and a compact decision space that allows the user to manage more than hundreds of attributes. A use case scenario of a prospective undergraduate student illustrated how annotations in DcPairs can be used to support the decision process. Second, the chapter examined a design-based debiasing method to alleviate the attraction effect. In particular, it presented a novel interaction technique based on a well-identified decision strategy (i.e. "elimination by aspects (EBA)"). An empirical study showed initial evidence that a manual deletion technique inspired by the "elimination by aspects" strategy can eliminate the attraction effect in scatterplot-based systems such as DcPairs.

7.2 Summary of Limitations

Specific limitations related to the empirical studies have been discussed in their individual chapters, and will not be repeated here. The goal of this section is not to be exhaustive, but to give a brief overview of the general limitations of the work.

Chapter 2 Chapter 2 presented an extensive review of decision theory and psychology literature. However, despite the systematic effort to present findings on human reasoning in a complete manner and derive accurate conclusions, it is possible that some important concepts are missing. Also, this dissertation was intended to be understood by people without a background in psychology or economics. As a result, some definitions of abstract concepts were rephrased in a simpler way to avoid the use of complex terminology. This approach carries the risk of loosing nuances of the original definition.

Another limitation of Chapter 2 relates to the FAULTY taxonomy of cognitive biases. First, in the textual description of the biases, examples of biases were omitted (although they are presented in Appendix A). Many biases are complex to explain and they need very elaborate explanations, that are beyond the scope of this dissertation. Moreover, it should be clear that the bias is indeed an error, but the violation of a norm is not always straightforward to grasp. For this purpose, the biases discussed were the ones whose violated norm felt easy to grasp, and there was no need for extensive explanations.

Another limitation of the taxonomy is that the analysis of the 139 papers presented (starting from 183 biases in wikipedia) has been conducted by a single coder (the author). Therefore, some biases

may have been mistakenly left out, and likely others are yet to be discovered. Also, the coder had to identify the user task in the procedure section of papers, which involves a fairly subjective judgment. Several independent coders need to validate the taxonomy.

Finally, there is a concern regarding the methodology of FAULTY. It is possible that the same cognitive bias exists in more than one task type in different academic papers. Even though most academic studies tend to consistently replicate the same tasks, this concern is indeed a possibility, but not necessarily a limitation. The assumption behind the classification of the FAULTY taxonomy is that different user tasks should be approached differently by researchers, even if the cognitive bias is currently known under the same umbrella term. A good example of such a case exists in the literature on the attraction effect. The attraction effect has been massively replicated as a task that involves a choice among three commercial products. However, some papers exist that tested the attraction effect in visual judgments, such as finding the largest rectangle [467] or finding similarities in circle and line pairs [101]. Even though these cases appear similar to the attraction effect and likely have similar roots, it is best if they are approached as perceptual biases, since people mainly fail to encode the visual property of an object.

Chapter 3 A limitation of Chapter 3 is that there was no sign of a clear difference in decision accuracy between the three techniques. One explanation is that the techniques are indeed comparable, but it is not clear if it is the case. Our decision accuracy metric showed a large variability, likely due to the fact that a multi-attribute choice task for holiday packages is inherently subjective. In addition, participants may not have been able to perfectly express (or be aware of) their criteria preferences, which likely adds further noise to the metric. As a result, this metric was not sensitive enough to capture differences that likely exist between conditions [106]. Additional work is needed to establish more sensitive metrics of decision quality, considering also non preference-based measures of success (e.g., statistical results in previous data in medical decisions).

Another limitation of Chapter 3 is the choice of the three techniques, PC, SM and TV. The systematic analysis that was conducted tried to identify the most suitable techniques for multi-attribute choice tasks. However, comparing three, likely comparable, candidates did not help us to calibrate our decision metric. Using also an inferior or a superior technique would have helped to understand the level of sensitivity of the metric. As a next step, other techniques that are specifically designed to support decisions would be evaluated. For example, Value Charts [85] and LineUp [201]combine tabular visualizations with stacked bar charts. The motivation behind not including these techniques was that elementary interactions need to be better understood before examining tools that combine techniques.

Chapter 4 Chapter 4 presented some negative results regarding the use of narratives in visualizations. However, researchers should not derive quick conclusions against the use of narratives in general. This experiment showed a rather artificial situation examining only short narratives in a crowdsourcing context. The main takeaway message from this chapter is that context should be used with care to be

successful, and that narratives may have complex and unanticipated effects, calling for more studies in this area.

Moreover, Chapter 4 did not attempt to explain how to design good narratives. Its goal was rather to answer the question: if a researcher adds a narrative when evaluating a visualization (as is done sometimes), should she expect performance to improve? This goal leaves room for imperfections in the wording of the narratives. More studies are however needed to understand the effect of narrative design, and whether better narratives exist that could be successful at improving job quality.

Although Chapter 4 found several clear effects (e.g., the accuracy drop caused by narratives, the lower performance of the decision-making framing for comparison tasks, and the negative subjective experience with abstract task framing) other effects are less conclusive, calling for follow-up studies. Chapter 4 uncovered what could be termed a "double-edged sword effect" of narratives, but does not provide detailed definitive explanations for all the effects observed. Future research will need to investigate why and how different types of narratives affect task performance and subjective experience. This research could involve, for example, interviewing crowdworkers. Finally, investigating the effect of narratives in lab settings would be another compelling route to explore.

Chapter 5 There are several limitations in Chapter 5. One stems from a general criticism of cognitive bias research, namely, that heuristics that appear irrational may not be so upon deeper examination [190]. Concerning the attraction effect, the way dominated alternatives are distributed could in some cases provide relevant information. For example, a real estate investor may infer from a region with many dominated alternatives that a certain type of house is more common, and therefore represents a larger market. At the same time, situations also exist where the number and position of dominated alternatives is clearly irrelevant and where a preference for the target would be irrational. This was the case for the "Bets" experiment in which the data does indicate that the vast majority of participants were unable to rationalize their choices based on where the dominated alternatives were located.

Another limitation of the Chapter 5 is that although it observed attraction effects, it did not investigate why they occur. In particular, it is unclear how much of the effect has cognitive vs. perceptual causes. Since in the "Bets" experiment regions with many decoys were visually more salient, it is possible that they drew participants' attention towards the target, or similarly, that participants sometimes failed to see the competitor because it was an isolated point. This possibility does not invalidate the existence of an attraction effect (as defined in Section 5.1), but it does raise the possibility that part of the effect with scatterplots (but not with numerical tables) has perceptual origins. The possible perceptual origins of the attraction effect were also analyzed in Section 6.3.1.

Chapter 6 The most important limitation in Chapter 6 is that the effectiveness of DcPairs has not been verified with a user study. For example, the assumption that the annotation feature can support the decision process has been only illustrated through a use case scenario (using LineUp's dataset [10, 201]). Moreover, although some basic elements of the design of DcPairs have been evaluated (such as the scatterplot matrix in Chapter 3, or the addition of a manual removal feature to alleviate

the attraction effect), it is unclear whether the overall system is effective. An evaluation methodology similar to the one presented in Chapter 3 could be used to compare DcPairs with a regular scatterplot matrix.

Another concern regarding the annotation feature of DcPairs is that it has been presented as a feature to add *data* into the system. However, adding data can have two meanings. The user adds new data cases or a new attribute. The difficulty in the second is that the system has to provide a representation for an "unexpected" attribute. DcPairs is focusing on adding a new attribute, but it needs further clarification. There is a question raised on whether adding a new attribute, e.g., some university locations [2] is data or metadata. As Munzner [347] admits, the distinction between data and metadata in the context of InfoVis is not clear. For example, Elias and Bezerianos [153] consider metadata as automatically extracted information (e.g. creation date). Ware [495] defines metadata as derived data after some operations on the original data that produced new insights, but he does not consider them as distinct from the data. In the context of annotating visualizations, we could think of metadata, to some extent, as data (about the original data) which make sense mostly for the user, and that the system likely ignores. For example, a simple textual annotation, that would not be captured by the system, could be better described as metadata. But by giving a concrete visual representation of this new data in the system (e.g., color), now the attribute makes sense both for the user but also for the system (since it is mapped in all other plots) and further interactions (e.g., filtering) are possible.

One final limitation of Chapter 6 is that the alleviation of the attraction effect was only observed in a single experiment, while the detection of the effect has been explored in a series of four experiments. Even though the results encourage the use of visualizations to alleviate cognitive biases, clearly more studies are needed to verify that manual deletion indeed alleviated the bias.

All chapters The crowdsourcing studies in Chapter 4, Chapter 5 and Chapter 6 used large samples (e.g. n=405 for narratives) to test a range of conditions and questions. However, their findings can be made more robust with additional studies testing alternative scenarios (e.g., other narratives, other choice sets for attraction effect), datasets, tasks, and performance metrics (such as open exploration and insight evaluation [412]). In addition, all studies focused on lossless geometric visualizations (scatterplots, tables, parallel coordinates), whereas more sophisticated visualizations should also be examined.

Due to the artificial nature of most of the studies, this dissertation mostly used synthetically generated datasets. This is a very common practice in decision making studies, but in the visualization field, the tested dataset can be an important factor. Thus, more realistic datasets need to be tested. However, this is not always straightforward. For example, the experiment "Real" in Chapter 5 showed inconclusive results with real datasets suggesting that the effects may be harder to measure there.

[2] Reminder of DcPairs: the university location was not part of the dataset, but could be added manually by the user in the form of colors (e.g., by marking with blue far-away universities)

7.3 Summary of contributions

Along with the empirical findings presented in Section 7.1, a list of contributions from this dissertation is outlined as follows (following the order of appearance in the chapters):

Chapter 2

- ✓ A review of common visualization tasks which concludes that decision making tasks are not supported by the current visualization taxonomies.
- ✓ An operational definition of a decision making task named *(multi-attribute) choice task* to help visualization researchers include decision tasks in their evaluations.
- ✓ A literature review of decision theory which presents how humans should, could and do make decisions, and outlines that decisions do not merely involve information understanding.
- ✓ A new taxonomy of 139 cognitive biases classified by user task, rather than by often untested explanations of why a bias occurs (as in previous taxonomies). This taxonomy can assist visualization designers to associate cognitive biases with visualization tasks, and provides pointers to the original methodologies for their detection. In particular, the choice class identifies 23 biases which could occur in *choice tasks*.
- ✓ A literature review of visualization works on cognitive biases showing that although many papers identified cognitive biases as important and studied them, there is little to no empirical evidence of biases occurring while using visualizations[3] and no validated work on alleviating them.
- ✓ A literature review of 38 decision-support visualization systems showing three main limitations: their evaluations lack decision metrics and sensible baselines of comparison; the tools provide limited interactions (no manual deletion, annotation, etc.) to aid decision strategies, and they misuse colors [4].

Chapter 3

- ✓ A systematic analysis of existing multidimensional visualizations which suggests that the family of "lossless" geometric visualizations is likely the most compatible with multi-attribute choice tasks.
- ✓ The first evaluation of multidimensional visualizations for their ability to support analytic and decision tasks, namely parallel coordinates (PC), scatterplot matrix (SM), and tabular visualization (TV), in which, despite the low attention they have received in the literature, TV showed compelling benefits.
- ✓ An implementation of PC, SM and TV that includes all standard features, keeps interactions and visual encodings consistent, and allows to present the same amount of information. The implementations are freely available online so that researchers can use them as effective baselines for other studies.
- ✓ A novel metric to assess decision quality based on repeated measures of self-reported preferences.
- ✓ A novel metric to assess decision confidence indirectly (attachment) which appeared to be more accurate than simple confidence ratings.

[3] One exception is a star-rating table from the HCI domain by Zhang et al. [530] (described in Section 2.3.5)

[4] The last two limitations are only briefly mentioned in Chapter 2 and mostly analysed in **Chapter 6**

✓ An example of a methodology to evaluate visualizations for decision-support that includes sensible baselines, consistent designs across visualizations, basic data exploration with analytic tasks, decision making tasks, and is primarily based on the outcome of the decision process.

Chapter 4

✓ The first information visualization study that investigates the effect of narratives in task instructions, showing subjective benefits for adding simple semantics to the dataset, but overall lower response accuracy for narratives.

Chapter 5

✓ A literature review on the *attraction effect* that highlights a debate in consumer research on whether the effect generalizes to non-numerical representations. The experiment "Gyms" suggests it does generalize. Psychologists also suggested that the effect should disappear, if more than three alternatives are added. The experiment "Bets" suggests it is not the case.

✓ An extension of the definition of the *attraction effect* to more than three alternatives and a procedure for constructing stimuli datasets.

✓ The first information visualization study showing that a cognitive bias called *attraction effect* can be present in visualizations, confirming that information can be well visualized and understood, and yet, the decision based on this information can be irrational.

Chapter 6

✓ The design and implementation of DcPairs, a decision-support visualization system that supports user-authored annotations, uncertainty communication, and a compact decision space that supports more than hundreds of attributes.

✓ A literature review and classification of debiasing methods, suggesting that extensive training in abstract rules shows some benefits, but re-designing the decision environment has limited empirical evidence for its effectiveness.

✓ The first information visualization study that provides initial evidence for an alleviation of a cognitive bias. The study showed that a manual deletion technique inspired by the "EBA" decision strategy could eliminate the attraction effect in scatterplot-based systems such as DcPairs.

7.4 Thesis Conclusion

As previously stated in Section 1.1, this dissertation makes a case for the following statement:

> ◇◇◇ To effectively support decision making, information visualization should move beyond the visual analysis paradigm. ◇◇◇

The visual analysis paradigm refers to the commonly accepted assumption in information visualization that the key purpose of visualization is to facilitate the understanding of information, such as the case of using a visualization tool to understand a complex dataset. Through extensive literature reviews and empirical studies, this dissertation suggests that the visual analysis paradigm cannot fully address the challenges of visualization-supported decision making.

Similarly to vision science that informs visualization researchers on the limitations of human vision, this dissertation reviewed decision theory to inform visualization researchers on the limitations of human reasoning. Decisions can be subjective, context-based, vulnerable to cognitive biases, and do not merely involve information understanding. Empirical findings from this dissertation suggest that these limitations persist even when the information is well-visualized and fully understood.

Decision tasks were neither part of visualization task taxonomies nor formally defined, and were usually omitted in the evaluations. A decision task has now been defined, and although it shares similar properties with analytic tasks, serves different user goals. The goal here is not to compare values, sort, determine ranges or correlations; the goal is to select the single best among several alternatives. Such decision tasks can be evaluated by using several metrics. Examples of such metrics are introduced in this dissertation: metrics based on preferences, dominance, confidence, as well as more sophisticated metrics that determine whether a user's decision is influenced by irrelevant information.

The evaluation methodologies of decision-support visualization systems are primarily based on data comprehension, such as whether users are able to use the system's interaction features, perform analytic tasks and gain insights. The evaluation methodology proposed by this dissertation suggests decision tasks as a complement, as well as sensible baselines and is primarily based on the outcome of the decision process. Evaluation methodologies can be further enriched by considering alternative well-established methodologies from the field of psychology. Pointers to such methodologies are now available thanks to a new task-based cognitive bias taxonomy.

Visual variables and representations are often designed with the criterion that the user can effectively process their perceptual properties. This dissertation also suggests that the cognitive properties of these variables, such as color semantics or whether they communicate decision uncertainty, should be used with care in visualization design. To effectively support decisions, visual representations ought to not distort judgement.

Most interactions offered by decision-support visualization systems do not focus on aiding different decision strategies. For example, users can not manually delete unwanted data. Moreover, although

decisions can be substantially based on external knowledge and subjective estimations, most interactions do not support the externalization of this knowledge on the visualization itself. For example, users cannot annotate data cases with additional attributes. This dissertation illustrates that enriching the palette of interactions offered by a system can better support the decision making process.

Current debiasing methods suggest that extensive training in abstract rules shows some benefits, but re-designing the decision environment has limited empirical evidence for its effectiveness. This dissertation suggests that a visualization tool that allows to clean up the decision space from unimportant or distracting information, can help users to focus on important information and make more rational decisions.

Visualization systems are mostly designed to assist professional analysts or data enthusiasts who are both primarily interested in understanding a dataset. This dissertation suggests moving beyond this "visual analysis paradigm" by focusing on another type of user: the decision maker. A decision maker can be anyone, from the manager of a company who needs to routinely make risky decisions to an ordinary person who wants to choose a career life path or simply find a camera to buy. The decision maker may or may not be interested in understanding a dataset, but still needs visualization to make data-informed and unbiased decisions.

Overall, this dissertation shows that moving beyond the visual analysis paradigm can contribute to making visualization a powerful decision support tool.

7.5 Future Perspectives

This section highlights opportunities for future research that stem from this work.

7.5.1 FAULTY: Cognitive bias exploration

The attraction effect (the black piece of the red bubble in Figure 7.5) examined in this dissertation, is only 1 of the 139 cognitive biases of the FAULTY taxonomy which can be explored in visualizations. This section attempts to illustrate the relevance of some of the other biases to visualization research.

The interest in cognitive biases shown by visualization community has grown a lot recently. One of the most prestigious visualization conferences, the IEEE Visualization Conference (VIS), runs a workshop explicitly dedicated to cognitive bias research [14], named "DECISIVe 2017" [4]. DECISIVe invited some very interesting visualization works-in-progress on cognitive biases.

In one of DECISIVe invited papers, Xiong et al. [519] provides preliminary but compelling evidence for the existence of another bias in visualizations: the *curse of knowledge* [81]. The curse of knowledge (■ #SEST) refers to experts who fail to foresee the difficulties of novices. In particular, Xiong et al. [519] showed that participants who read a narrative related to the data consider as more salient the patterns that are compatible with this narrative. Furthermore, after acquiring this background knowledge from the narrative, participants predict that other viewers would consider the same patterns as more salient, as well. This work emphasizes that failing to have an accurate idea of the knowledge of novice users

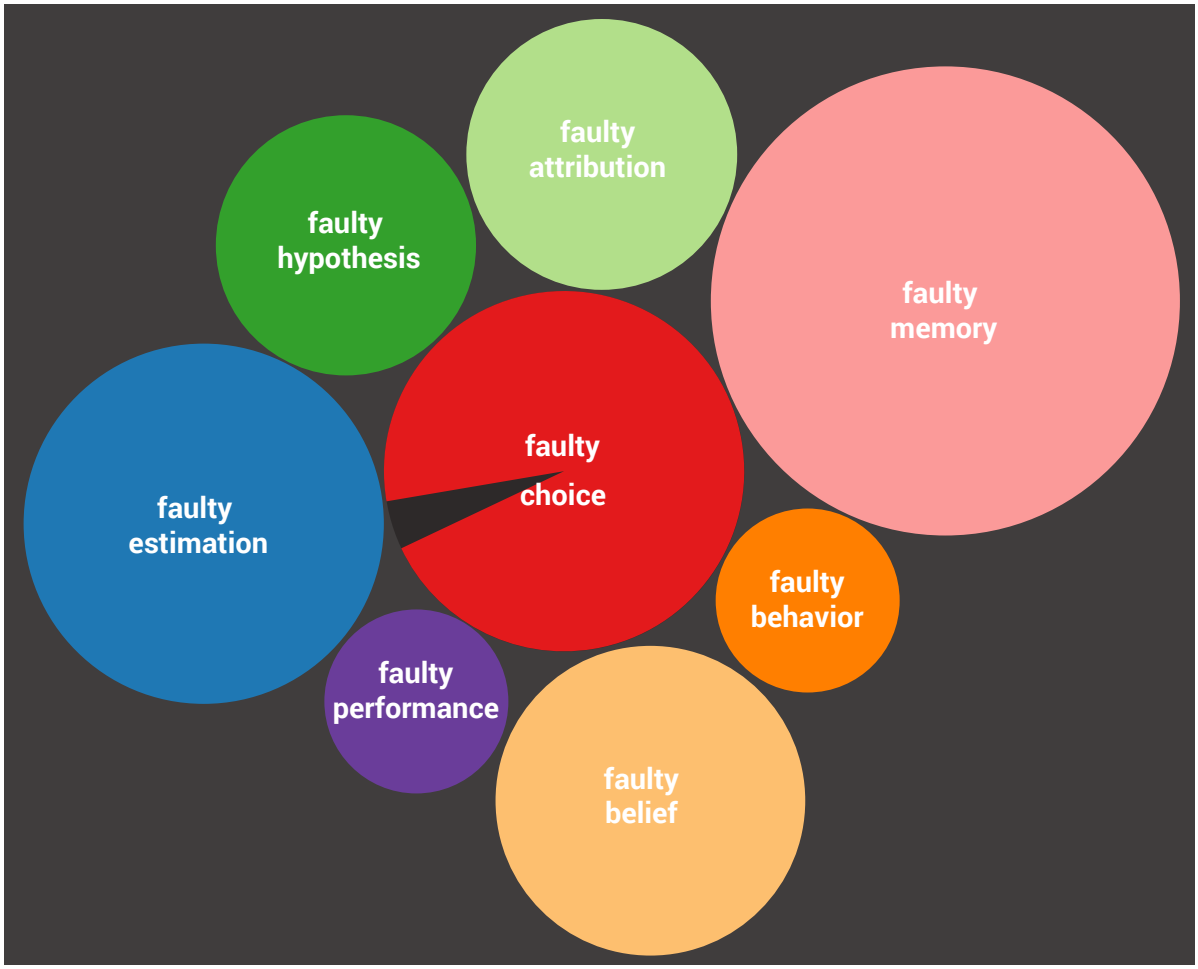


Figure 7.5: FAULTY: taxonomy of cognitive biases.

can have important implications in visual data communication. Furthermore, along with the results of the narratives experiment in Chapter 4, these results illustrate that narratives should rather be used with care with data visualizations. In another DECISIVe invited paper, though, Pohl [381] illustrates a different approach regarding narratives and visualization biases. Pohl emphasizes that many tasks that are typically tested in cognitive bias studies are context-free and that there are cases where the reason for the bias is simply a lack of knowledge. The author suggests instead that visualization designs should provide more context to activate background knowledge [381]. In particular, Pohl [381] discusses the use case of crime analysts who often report that in order to make a judgment, they need more detailed visualizations than abstract node-link representations.

Other recent works attempt to build a common vocabulary [494] and frameworks [493] to help visualization researchers understand how to approach cognitive biases. To do this, most papers rely on explanations of psychology research of why each bias occurs, e.g., people's tendency to for quick, automated decisions rather than deliberate ones or because of information overload [494]. In the same line, Wall et al. discussed six possible metrics (Figure 7.6) to detect and quantify biases in visualization systems [493]. Even though these metrics can be extremely useful for quantitative evaluations, rely on the inspection of whether the user has explored a subset or the overall dataset. This approach follows a visual analysis paradigm, as defined in this dissertation, according to which access and comprehension of more information will necessarily lead to better decisions. However, this dissertation illustrated that this assumption is not always correct. Therefore, the task-based FAULTY taxonomy can further indicate possible visualization tasks that are prone to biases.

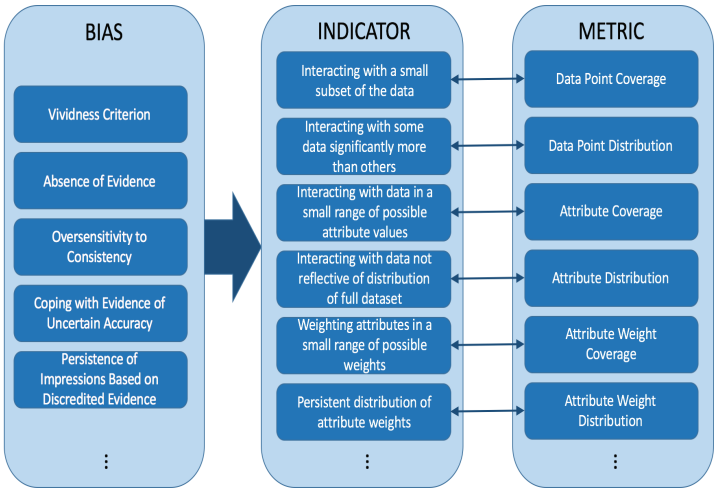


Figure 7.6: Six bias metrics by Wall et al.[493]

Considering possible extensions to the attraction effect findings, most of most biases of a faulty choice category (■ #CHOI) are likely to affect choice tasks with visualizations. For example, the *compromise effect* is a similar to the attraction effect in that people tend to choose the middle in a given set of alternatives [433]. Another example is the *phantom effect* in which people's choices are affected by dominant but unavailable alternatives [376]. All these biases, like in the attraction effect, have been defined and tested in 3 alternatives using non-visual formats. Therefore their possible extension to larger datasets is not straightforward to apply (a possible illustration in shown in Figure 7.7). Another challenge here is that once more alternatives are added, these biases interfere with each other and the effects may be harder to measure.

Another important family of biases involves decision making under uncertainty. The biases of a faulty estimation task (■, #SEST, #PROB) illustrate that people very often experience problems in understanding the notion of randomness. As a result, they may see random patterns such as streaks, or clusters in large samples of random data. Visualization research has already examined the problem of communicating randomness when analysts are making statistical inferences [508]. An effective visualization design should ideally be able to communicate the level of randomness in the data. For example, in a standard scatterplot (Figure 7.8), it may be hard to recognize which dataset is randomly generated. Even though these questions seem critical for visualized decision-making and analysis, it seems that there is no empirical work in visualizations which studies such questions, for example, which visualization designs are more effective in communicating randomness in the data.

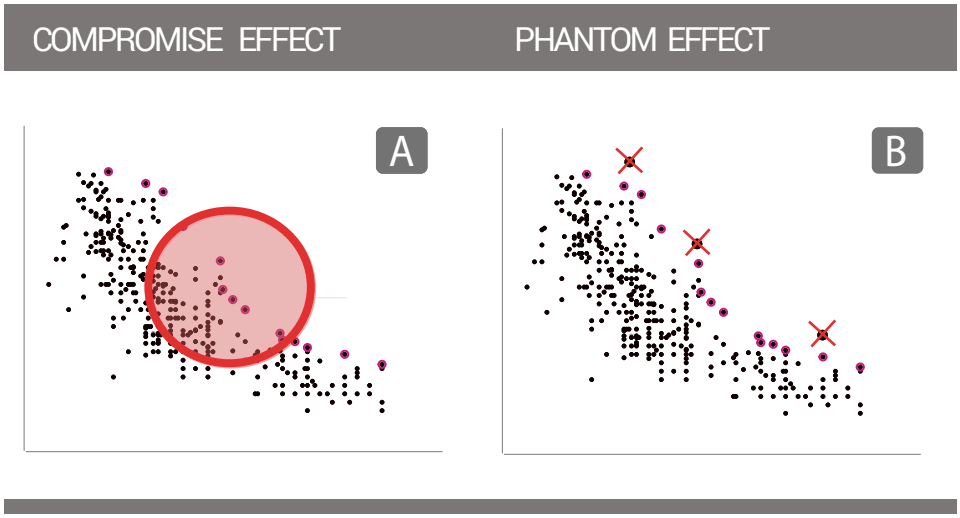


Figure 7.7: Possible extensions to the attraction effect: *compromise effect* (A), where peoples' choices are affected by the middle alternatives, and *phantom effect* (B), by dominant but unavailable alternatives.

Many other visualization topics can be related to cognitive biases. For example, there are visualization works that associate the memorability of a visualization with its effectiveness [62, 215]. It seems that there are no previous studies investigating which visualizations can be more vulnerable to memory biases. Such memory limitations also have implications in several other tasks besides recall. An example is how serial position memory biases [348] can affect choice tasks (e.g. selecting candidates according to the position of their name in the ballot paper [516]), or perceptual tasks [481]. Another recent trend in information visualization is progressive analytics [448]. When data becomes large, interactive exploration can be severely hampered. As a result, analysts have to monitor the progression of the results, and the visualization designer has to account for progressive representations of the data. Many biases in the FAULTY taxonomy are related to the way people remain conservative in the light of new information [147, 149]. Such faulty tasks could be used to evaluate whether a progressive design effectively communicates the representation updates.

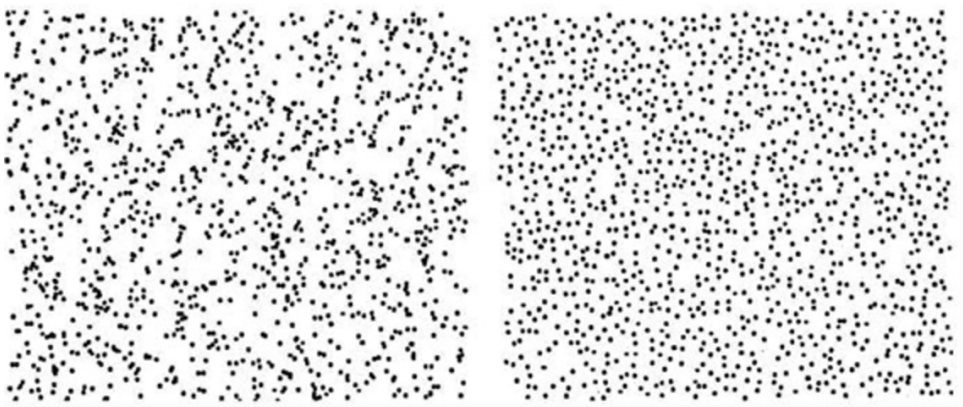


Figure 7.8: Which plot was randomly generated? Answer: the left, source [2]



Figure 7.9: Group decision making with visualizations. source: A webinar led by Tara Nutley, MEASURE Evaluation; Stacey Berlow, Project

7.5.2 Group decision making

This dissertation showed that the process of decision making is often suboptimal for single individuals even when facing relatively simple situations. For example, in the first attraction effect experiment, people were irrational when choosing between a clean gym and a gym with a large variety of machines. Decisions involving large datasets are likely to be more challenging and subject to more biases, since large datasets escape the understanding of even domain experts. Although collaboration (Figure 7.9) has been shown to yield better decisions in some cases [453], involving groups adds another level of complexity to the problem: group decisions can involve large numbers of people [32], participants may have conflicting interests [80, 365], and some people may exert a detrimental influence [34].

In order to alleviate cognitive biases, this dissertation explored ways to enrich a visualization system with new interaction techniques. As discussed in Section 6.2, group decision making is another promising debiasing approach, that introduces new challenges, designs and interactions to explore. The most effective way to make a group decision is believed to be when individuals decide independently and the final decision is aggregated [232], so that suboptimal strategies can be eliminated [438]. When group members form their judgments during discussion though, individual biases can be amplified [380]. Group interactions are also known to trigger other cognitive biases such as shaping favorable judgments for the group one belongs to [79, 377] and they are often biased towards conformity [169] or polarization [344]. However, there is some recent contradicting evidence that deliberation and discussion can improve collective wisdom [351, 385] depending on how the members of the group



Figure 7.10: Automation bias source: lskitka.people.uic.edu/styled-7/styled-14/

interact with each other. Along the same lines, the use of online rating data to help make informed decisions, e.g., to decide which digital camera to buy, has been investigated [300], and there is evidence that financial choices of a social network can help older adults to overcome their biases in financial risk taking [531].

These different issues have already been widely studied (e.g., in economics, game theory, voting theory, ethical philosophy and social psychology) and the "wisdom of crowds" is already used to solve problems such as diagnosis of cancer and financial forecasting [351]. However, these topics are not yet examined in combination with visual analytics support tools. Moreover, in addition to the general lack of understanding of the dynamics of group decision making involving large datasets, there is a lack of effective and generic visual tools for supporting such decisions.

7.5.3 Computation-aided decision making

This dissertation examined multi-attribute choice tasks that involve personal preferences. However, as discussed in their definition, multi-attribute choice tasks can also involve cases where choice goodness is defined objectively in terms of prior evidence. For example, doctors may need to make a data-driven decision that is based on a large number of attributes and results from several research studies. In very complex decisions, computations of statistical models based on past data can be helpful for the users

[129]. For example, in clinical judgments that systematically suffer from judgment inconsistencies [129], automated decision support systems (DSS) can run consistency checks (e.g., on attribute weights or probabilities) [256]. Moreover, computational tools can incorporate normative algorithms into the decision making process which would be otherwise too difficult, if not impossible, for human beings to compute [124]. Therefore, the assistance of automated analysis can be also critical in complex decisions.

Automated analysis tools range from simple statistical procedures, to expert systems involving machine learning, artificial intelligence, and data mining. Although these approaches are reaching maturity, they are not yet able to support complex decisions for several reasons. First, they ignore expert knowledge that cannot be formalized [97, 464, 503]. Then, analysis tools often act as a "black box" that are hard to predict and understand, sometimes even by experts [186]. In other cases, analysis tools restrict users by expecting a very particular input, ignoring other context-relevant information that the users may have [459]. Thus automatic analysis is powerful, but often results in tools that are difficult to control and predict, and that leave little room for subjective judgment. In contrast with automatic analysis, information visualization delegates most of the thinking to humans by capitalizing on visual perception [84, 160]. Decision making in complex problems is likely most effective when remaining under the control of end users, while also being assisted with computer analysis that is informed by user expertise [90].

A challenging new route to explore would be to investigate visualization decision-support systems that combine the strengths of humans with the strengths of automatic analysis. An interesting use case of challenges arising in such a combination is illustrated by a type of faulty choice bias (■ #CHOI), the *automation bias*. When certain information is given by an automated system, people tend to over-rely on it even when their expertise and real-world evidence suggest otherwise [124]. For example, pilots often fail to make the right decision over erroneous autopilot recommendations [124] (Figure 7.10). Visualizations provide an opportunity to help deal with challenges in combining human judgment with automatic analysis, for example, to investigate designs that effectively communicate uncertainty over automated suggestions [21, 450].

7.5.4 Final words

Overall, the limitations of human reasoning can often be challenging for visualization researchers to explore. However, the prospect of building systems that can empower human decisions is beyond doubt intriguing and rewarding. The *rhyme-as-reason* [330] bias suggests that conclusions are perceived as more truthful once rewritten to rhyme. Therefore, the final words of this dissertation will be left to C.P. CAVAFY on the page that follows.

ITHACA

As you set out bound for Ithaca,
hope that the journey is a long one,
full of adventures, full of learning.
Of the Laestrygonians and Cyclopes,
of wrathful Poseidon have no fear,
you'll never meet suchlike on your journey,
if your thoughts remain lofty, if noble
sentiment grips your body and spirit.
You'll never encounter raging Poseidon,
Laestrygonians and Cyclopes,
unless you bear them in your soul,
unless your soul sets them before you.

Hope that the journey is a long one.
That the summer morns be many
when with what delight, what joy
you enter harbours hitherto unseen;
that you stop at Phoenician markets
and acquire fine merchandise,
nacre and coral, amber and ebony,
and all kinds of heady perfumes,
as many heady perfumes as you can;
that you visit many Egyptian cities,
to learn and learn from the erudite.

Always keep Ithaca in mind.
To arrive there is your destination.
But in no way rush the voyage.
Better for it to last many years;
and for you to berth on the isle an old man,
rich with all you gained on the journey,
without expecting Ithaca to give you riches.

Ithaca gave you the wonderful voyage.
Without her you would not have set out on your way.
Yet she has nothing more to give you.

And though you may find her wanting, Ithaca has not
deceive you.

Wise as you've become, with so much experience,
already you'll have understood what these Ithacas mean.

C.P. CAVAFY translated by David Connolly. ΑΙΩΡΑ



APPENDIX A

A.1 FAULTY taxonomy of cognitive biases

The column "Category" in Table A.1 indicates the category of each bias as shown in Table A.1. The column "Cognitive bias" in Table A.1 reports the name of each bias. Synonym names of cognitive biases are reported in Table A.3. The column "REF" in Table A.1 is a peer-reviewed academic paper,

Table A.1: Color legend of the task-based categories of cognitive biases

1	#CHOI	Biases of a faulty choice task
2	#ESTI	Biases of a faulty estimation task (quantitative or qualitative estimation)
3	#MEMO	Biases of a faulty recall task
4	#HYPO	Biases of a faulty hypothesis assessment task
5	#ATTR	Biases of a faulty attribution task
6	#PERF	Biases of a faulty performance evaluation task (after a given puzzle)
7	#BELI	Biases of a faulty belief task (e.g., moral,social,political, personality traits)
8	#BEHA	Biases of a faulty behavior (no instructed task, behavior observation)

Table A.2: Color legend of the relation of each cognitive bias to visualization research

#8	evidence for the alleviation of the cognitive bias in visualization
#7	evidence for the detection of the cognitive bias in visualization
#6	studied in the context of visualizations, but yet unclear if any of the above
#5	discussed in visualization research as important, but not yet studied
#4	not discussed in visualization but likely relevant
#3	probably relevant to visualization
#2	potentially relevant to visualization
#1	relevance to visualization currently unclear

that either contains an experiment that tests the bias for humans, or describes the experiment citing the original paper.

The column "InfoVis" in Table A.1 indicates whether each cognitive bias has been examined in information visualization research and reports the reference of the paper. As shown in Table A.2, cognitive biases which have been alleviated on visualizations are marked with #8. Cognitive biases which have been detected on visualizations are marked with #7. Cognitive biases which have been studied in the context of visualizations, but without evidence of detection or alleviation of the bias, are marked with #6. Cognitive biases which have been discussed in visualization research as important, but have not yet been studied, are marked with #5. Many of these biases are described in Chapter 2. The remaining marks #4, #3, #2 and #1 indicate some interesting cognitive biases to study in an information visualization context.

#	Category	Cognitive Bias	REF	InfoVis	Short Explanation
1	#ESTI	Anchoring effect	[181]	#7 [98, 481]	estimation affected by first piece of information
2	#ESTI	Availability bias	[473]	#5 [114, 132, 536]	events more probable if easy to remember
3	#ESTI	Base rate fallacy	[41]	#6 [276, 332]	ignore base rate probability of general population
4	#ESTI	Conjunction fallacy	[476]	#5 [536]	specific events more probable than general
5	#ESTI	Conservatism	[378]	#7 [530]	adjust probability estimation insufficiently in light of new information
6	#ESTI	Curse of knowledge	[275]	#7 [519]	experts fail to predict the judgments of novices
7	#ESTI	Empathy gap	[308]	#1	ability to predict future behavior affected by current emotions
8	#ESTI	Exaggerated expectation	[490]	#4	real-world evidence less extreme than expected
9	#ESTI	Extrinsic incentives bias	[216]	#1	predict extrinsic motivations for others (e.g.money) intrinsic to oneself (e.g.learnin
10	#ESTI	False consensus effect	[400]	#2	overestimate the degree to which others agree with oneself
11	#ESTI	Gambler's fallacy	[474]	#4	an event currently more frequent than normal will be less frequent in the future
12	#ESTI	Hot-hand fallacy	[195]	#5 [114]	more probable successful attempt if previous attempt was successful
13	#ESTI	Illusion of control	[462]	#3	overestimation of one's influence on an external event
14	#ESTI	Impact bias	[410]	#1	predict future emotional reactions as more intense
15	#ESTI	Insensitivity to sample size	[474]	#5 [154, 536]	estimate probability ignoring sample size
16	#ESTI	Naive cynicism	[289]	#2	predict that the others will be more egocentrically biased
17	#ESTI	Optimism bias	[505]	#4	positive outcomes more probable for oneself than others
18	#ESTI	Out-group homogeneity bias	[367]	#4	estimate out-group will be more homogenous than in-group members
19	#ESTI	Pessimism bias	[425]	#4	positive outcomes less probable for oneself than others
20	#ESTI	Planning fallacy	[78]	#5 [141]	overoptimistic completion time (faster for oneself than for others)
21	#ESTI	Regressive bias	[36]	#4	overestimate high probabilities, underestimate low ones
22	#ESTI	Restraint bias	[354]	#1	overestimate one's ability to resist temptation
23	#ESTI	Sexual over/under-perception bias	[212]	#1	over-/underestimate probability of sexual interest of others
24	#ESTI	Spotlight effect	[196]	#1	overestimate probability that people notice one's appearance/behavior
25	#ESTI	Subadditivity effect	[478]	#4	overall probability less than the probabilities of the parts
26	#ESTI	Time-saving bias	[455]	#4	over/underestimate time saved/lost when increasing/decreasing speed
27	#ESTI	Weber-Fechner law	[100]	#6 [?]	fail to estimate small differences in large quantities
28	#CHOI	Ambiguity effect	[397]	#4	choices affected by their association with unknown outcomes
29	#CHOI	Attraction effect	[240]	#8 [134]	choices affected by irrelevant dominated alternatives
30	#CHOI	Automation bias	[317]	#5 [407]	choices affected by automated system recommendations
31	#CHOI	Ballot names bias	[516]	#6 [516]	voting choices affected by the order of candidate names
32	#CHOI	Cheerleader effect	[491]	#1	people are more attractive when in a group
33	#CHOI	Compromise effect	[433]	#5 [134]	choices affected if presented as extreme or average alternatives
34	#CHOI	Denomination effect	[390]	#4	choices affected by whether item corresponds to large amount or small multiple or
35	#CHOI	Disposition effect	[499]	#4	selling choices affected by initial and not current value
36	#CHOI	Distinction bias	[236]	#4	choices affected by whether items are evaluated simultaneously or separately
37	#CHOI	Endowment effect	[340]	#4	choices affected by ownership of alternatives
38	#CHOI	Escalation of commitment	[447]	#4	choices affected by previously allocated resources despite negative outcome
39	#CHOI	Framing effect	[475]	#5 [154, 536]	choices affected if presented as gains or losses
40	#CHOI	Hyperbolic discounting	[460]	#4	choices affected by smaller, but short-term rewards

#	Category	Cognitive Bias	REF	InfoVis	Short Explanation
41	#CHOI	Identifiable victim effect	[255]	#6 [67]	donation choices affected by whether victims are identifiable
42	#CHOI	IKEA effect	[355]	#5 [449]	choices affected by whether alternatives involved self-effort
43	#CHOI	Less is better effect	[237]	#4	choices affected if presented separately or juxtaposed
44	#CHOI	Loss aversion	[461]	#7 [128, 530]	choices affected by whether alternatives are gains or losses
45	#CHOI	Mere-exposure effect	[527]	#5 [154, 268]	choices affected by familiarity (repeated exposure)
46	#CHOI	Money illusion	[423]	#4	choices affected by nominal monetary values
47	#CHOI	Neglect of probability	[452]	#4	completely disregard probabilities in uncertain choices
48	#CHOI	Not invented here	[26]	#4	choices affected by whether alternatives are of external origin
49	#CHOI	Phantom effect	[376]	#4	choices affected by dominant but unavailable alternatives
50	#CHOI	Pseudocertainty effect	[477]	#5 [536]	choices affected by whether some alternatives are framed as certain
51	#CHOI	Reactance	[69]	#4	choices affected by the urge to choose the opposite of what someone wants you to do
52	#CHOI	Reactive devaluation	[399]	#4	choices affected if allegedly originated with an antagonist
53	#CHOI	Shared information bias	[40]	#4	group choices affected by members not sharing new information
54	#CHOI	Status quo bias	[408]	#5 [360]	choices affected by whether to keep current state or require change
55	#CHOI	Social comparison bias	[183]	#1	hiring choices affected by avoiding similar competences
56	#CHOI	Well traveled road effect	[250]	#4	travel road choices affected by road familiarity
57	#CHOI	Zero-risk bias	[44]	#4	choices affected by alternatives with complete risk elimination
58	#HYPO	Barnum effect	[463]	#2	high accuracy ratings for vague and general personality statements
59	#HYPO	Belief bias	[158]	#3	hypothesis true if conclusion is believable
60	#HYPO	Clustering illusion	[264]	#5 [154?]	consider patterns as non-random in small samples to confirm a hypothesis
61	#HYPO	Confirmation bias	[315]	#5 [154]	favor any strategy (e.g. reasoning, info seeking) that confirms initial hypothesis
62	#HYPO	Congruence bias	[496]	#4	test only direct hypothesis
63	#HYPO	Experimenter effect	[398]	#3	subconsciously influence responses to confirm a hypothesis
64	#HYPO	Illusory correlation	[89]	#5 [536]	perceived relationship between variables that does not exist
65	#HYPO	Illusory truth effect	[213]	#4	statement considered true after repeated exposure to it
66	#HYPO	Information bias	[43]	#4	seek extra information irrelevant to the hypothesis
67	#HYPO	Pareidolia	[489]	#4	consider a random pattern as meaningful
68	#HYPO	Rhyme as reason effect	[330]	#1	statement true if framed as a rhyme
69	#ATTR	Actor-observer asymmetry	[335]	#3	one's own failures attributed to situation; others' to personality weaknesses
70	#ATTR	Defensive attribution hypothesis	[426]	#1	attribution affected by one's similarity with the actor
71	#ATTR	Egocentric bias	[401]	#3	the result of a joined action attributed mostly to oneself than to others
72	#ATTR	Fundamental attribution error	[193]	#1	others' actions attributed to their character rather than the situation
73	#ATTR	Group attribution error	[18]	#2	group actions attributed to an individual member
74	#ATTR	Hostile attribution bias	[200]	#1	ambiguous behavior attributed to intentional incentives
75	#ATTR	Illusion of external agency	[194]	#2	satisfactory outcomes attributed to external agents
76	#ATTR	In-group favoritism	[458]	#1	success attributed to in-group members over out-group
77	#ATTR	Just-world hypothesis	[206]	#2	consequences attributed to the assumption that the world is fundamentally just
78	#ATTR	Self-serving bias	[83]	#2	one's own achievement attributed to ability/effort; failure to external factors
79	#ATTR	System justification	[261]	#3	status quo attributed to the system
80	#ATTR	Trait ascription bias	[387]	#2	attribute varied traits to oneself, homogenous to others
81	#ATTR	Ultimate attribution error	[377]	#3	favor in-group and negatively attribute out-groups (like Actor-observer asymmetry)
82	#MEMO	Bizarreness effect	[326]	#4	easier to recall bizarre items
83	#MEMO	Childhood amnesia	[480]	#1	harder to recall events (e.g.times, places, emotions, people) before certain age
84	#MEMO	Choice-supportive bias	[322]	#2	recall past choices as better than they were
85	#MEMO	Continued influence effect	[259]	#5 [493]	misinformation affects beliefs even after correction
86	#MEMO	Cross-race effect	[457]	#1	hard to recall people of different race
87	#MEMO	Cryptomnesia	[75]	#3	memory mistaken for imagination, inspiration (e.g. unintentional plagiarism)
88	#MEMO	Cue-dependent forgetting	[470]	#4	failure to recall information without memory cues
89	#MEMO	Digital amnesia	[440]	#3	harder to recall information that can be found in a search engine
90	#MEMO	Duration neglect	[175]	#2	recall unpleasant experiences according to intensity, ignoring duration
91	#MEMO	Fading affect bias	[492]	#1	forget emotion of unpleasant events, but recall pleasant
92	#MEMO	False memory	[309]	#1	imagination mistaken for a memory
93	#MEMO	Hindsight bias	[165]	#5 [536]	recall past predictions as more accurate after seeing the outcome
94	#MEMO	Humor effect	[451]	#3	easier to recall humorous items
95	#MEMO	Leveling and sharpening	[285]	#4	on recall exaggeration of selected features (sharpening)/weakening others (leveling)
96	#MEMO	Levels-of-processing effect	[115]	#4	easier to recall result of deep level analysis
97	#MEMO	List-length effect	[88]	#4	harder to recall items from longer lists
98	#MEMO	Misinformation effect	[370]	#3	misinformation of events results in a false recall

#	Category	Cognitive Bias	REF	InfoVis	Short Explanation
99	#MEMO	Modality effect	[187]	#4	easier to recall items presented auditorily than visually
100	#MEMO	Mood-congruent memory	[324]	#1	recall affected by current mood
101	#MEMO	Next-in-line effect	[72]	#2	failure to recall words of previous speaker in turns speaking
102	#MEMO	Part-list cueing effect	[436]	#4	harder to recall material after reexposure to subset
103	#MEMO	Picture superiority effect	[325]	#4	easier to recall images (symbolic representations) than words
104	#MEMO	Positivity effect	[321]	#1	easier to recall positive events than negative (stronger in elderly people)
105	#MEMO	Processing difficulty effect	[357]	#4	easier to recall information which was hard to comprehend
106	#MEMO	Reminiscence bump	[252]	#1	easier to recall events from adolescence and early adulthood
107	#MEMO	Rosy retrospection	[338]	#1	remember past better than it was
108	#MEMO	Self-generation effect	[120]	#4	self-generated content is easier to recall than if simply read
109	#MEMO	Self-reference effect	[?]	#1	easier to recall self-related information
110	#MEMO	Serial-positioning effect	[348]	#4	best recall first (primacy) and last (recency) items in a series
111	#MEMO	Source confusion	[225]	#2	memory distorted after hearing people speak about a situation
112	#MEMO	Spacing effect	[204]	#3	easier to recall information repeated in widely spaced times than massed
113	#MEMO	Suffix effect	[343]	#2	recency effect diminished by an irrelevant suffix sound at the end of a list
114	#MEMO	Suggestibility	[306]	#1	ideas suggested by a questioner mistaken for memory
115	#MEMO	Telescoping effect	[253]	#3	temporal displacement of an event
116	#MEMO	Testing effect	[328]	#3	easier to recall information tested (classroom style test)
117	#MEMO	Tip of the tongue phenomenon	[74]	#1	recall parts of an item but not the whole
118	#MEMO	Verbatim effect	[383]	#1	easier to recall a “gist” than a verbatim wording
119	#MEMO	Von Restorff effect	[368]	#4	easier to recall a distinct item among a list of similar items
120	#MEMO	Zeigarnik effect	[152]	#1	easier to recall interrupted tasks than completed
121	#PERF	Dunning-Kruger effect	[288]	#5 [?]	low-ability people overestimate their performance (opposite for high-ability)
122	#PERF	Hard-easy effect	[304]	#3	overconfidence for hard tasks, underconfidence for easy
123	#PERF	Illusion of transparency	[416]	#1	evaluate performance in public more harshly (e.g. visible speech anxiety)
124	#PERF	Illusion of validity	[474]	#5 [536]	overconfidence for highly fallible predictions if story is coherent
125	#PERF	Outcome bias	[42]	#2	evaluate decision maker only by choice outcome
126	#PERF	Worse-than-average effect	[287]	#3	underestimate one’s achievements related to others in difficult tasks
127	#BELI	Anthropocentric thinking	[107]	#1	humans as a default analogical base for reasoning about biological species
128	#BELI	Anthropomorphism	[498]	#5 [67]	belief that non-human environments operate similarly to a human’s
129	#BELI	Backfire effect	[356]	#3	prior beliefs stronger when corrected
130	#BELI	Bandwagon effect	[350]	#2	beliefs affected by opinions of others
131	#BELI	Ben Franklin effect	[254]	#1	opinion of others is affected by one’s behavior towards them
132	#BELI	Bias blind spot	[386]	#2	belief that biases are more prevalent in others than oneself
133	#BELI	Focusing effect	[417]	#4	beliefs based on the most pronounced part of given information
134	#BELI	Halo effect	[353]	#2	personality trait ascription affected by overall attractiveness
135	#BELI	Illusion of asymmetric insight	[388]	#1	belief that one knows more about others than others know about oneself
136	#BELI	Illusory superiority	[145]	#2	personality traits favorable to oneself over others
137	#BELI	Moral credential effect	[339]	#1	more prejudiced beliefs once receive non-prejudice credentials
138	#BELI	Moral luck	[125]	#1	ascribe more or less moral blame based on outcome of an event
139	#BELI	Naive realism	[214]	#1	belief we experience objects in our world objectively
140	#BELI	Negativity bias	[170]	#3	social judgments affected more by negative than positive information
141	#BELI	Omission bias	[444]	#3	moral judgments affected by whether the harm was caused by inaction
142	#BELI	Social desirability bias	[119]	#2	respond in questionnaires in a socially approved manner
143	#BELI	Third-person effect	[25]	#2	others more vulnerable to mass media messages than oneself
144	#BELI	Women are wonderful effect	[404]	#1	associate more positive characteristics to women
145	#BELI	Zero-sum bias	[403]	#1	belief that one’s gain is another one’s loss
146	#BEHA	Attentional bias	[251]	#3	current specific thoughts affect attention
147	#BEHA	Authority bias	[334]	#1	tendency to obey an authoritarian figure
148	#BEHA	Functional fixedness	[185]	#4	tendency to use objects only in the traditional way
149	#BEHA	Ostrich effect	[269]	#4	investors avoid monitoring portfolios to ignore negative information
150	#BEHA	Risk compensation	[220]	#3	take greater risks when perceived safety increases
151	#BEHA	Semmelweis reflex	[342]	#3	medical practices follow old established norms and do not adapt to new knowledge
152	#BEHA	Stereotyping	[113]	#3	behave to people in a way that expects certain characteristics without prior info
153	#BEHA	Surrogation	[99]	#3	managers behave as if measures that represent constructs are actually the constructs
154	#BEHA	Unit bias	[184]	#1	eat more food in bigger containers

Bias	Synonym / Similar Biases
Actor-observer asymmetry	Actor-observer bias
Anchoring effect	Focalism or Anchoring
Anthropomorphism	Personification
Attraction effect	Decoy effect or Asymmetric dominance effect
Availability bias	Availability heuristic
Base rate fallacy	Base rate neglect or Base rate bias or Extension neglect
Childhood amnesia	Infantile amnesia
Choice-supportive bias	Post-purchase rationalization
Clustering illusion	Texas sharpshooter fallacy
Confirmation bias	Confirmatory bias or Myside bias or Semmelweis reflex
Conjunction fallacy	Linda problem
Conservatism	Belief revision
Cross-race effect	Cross-race bias or Other-race bias or Own-race bias
Cryptomnesia	Inadvertent plagiarism
Cue-dependent forgetting	Context effect or Retrieval failure
Digital amnesia	Google effect
Duration neglect	Extension neglect or Peak-end rule
Hard-easy effect	Discriminability effect or Difficulty effect
Hyperbolic discounting	Dynamic inconsistency
Empathy gap	Projection bias
Endowment effect	Divestiture aversion or Mere ownership effect
Escalation of commitment	Irrational escalation of commitment or Commitment bias or Sunk cost fallacy
Experimenter effect	Experimenter-expectancy effect or Observer-expectancy effect or Experimenter's bias or Expectancy bias or Subject-expectancy effect
Fading affect bias	FAD
Focusing effect	Focusing illusion
Barnum effect	Forer effect or Subjective validation or Acceptance phenomenon or Personal validation
Fundamental attribution error	Correspondence bias or Attribution effect
Gambler's fallacy	Monte Carlo fallacy or Fallacy of the maturity of chances
Illusory truth effect	Truth effect or Illusion-of-truth effect or Reiteration effect or Validity effect or Frequency-validity relationship or Availability cascade
Illusory superiority	Above-average effect or Superiority bias or Leniency error or Sense of relative superiority or Primus inter pares effect or Lake Wobegon effect
Impact bias	Durability bias
In-group favoritism	In-group bias or In-group/out-group bias or Intergroup bias
Hindsight bias	Knew-it-all-along effect or Creeping determinism
Hostile attribution bias	Hostile attribution of intent
Hot-hand fallacy	Hot hand phenomenon or Hot hand
IKEA effect	Effort justification
Lag effect	Spacing effect or Maslow's gavel or Golden hammer
Mere exposure effect	Familiarity principle
Money illusion	Price illusion
Negativity bias	Negativity effect
Not invented here	NIH
Optimism bias	Unrealistic optimism or Comparative optimism
Pessimism bias	Pessimistic bias
Rhyme-as-reason effect	Eaton-Rosen phenomenon
Risk compensation	Peltzman effect
Self-generation effect	Generation effect
Self-reference effect	Self-relevance effect
Serial position effect	Recency effect or Primacy effect
Survivorship bias	Survival bias
Telescoping effect	Telescoping bias
Testing effect	Retrieval practice or Practice testing or Test-enhanced learning
Triviality	Parkinson's Law of
Von Restorff effect	Isolation effect
Weber-Fechner law	Weber's law or Fechner's law
Worse-than-average effect	Below-average effect

Table A.3: Alternative names for cognitive biases

APPENDIX B

Appendix B

B.1 Decision-Support Visualization Systems

Table B.1: Color Legend for Decision Support Visualization Systems

I = **Independent**

■ generic tools that can support any data table and multi-attribue tasks

– domain specific tools that their decign can be used with the specific domain dataset

B = **Baseline**

■ compared to another visualization system

■ compared to a non visualization system (e.g. static table, webform, database, text)
or to a design variations of the same tool

– no baseline

E = **Evaluation**

#1 quantitative with decision tasks, with decision accuracy validation
#2 quantitative with decision tasks
#3 quantitative with analytic tasks
#4 qualitative with decision tasks
#5 qualitative
– no user evaluation

Table B.2: Decision Support Visualization Systems Index

#	REF	I	B	E	Name	Short Description
1	[454]			#5	ADVISES	epidemiology decision support
2	[15]				Afzal et al.	decision support for epidemic modeling
3	[33]			#5	AHP treemaps	treemaps to visualize the analytic hierarchy process
4	[442]				Attribute Explorer	interactive linked histograms
5	[301]				Classimap	mapping technique for decision support
6	[23, 24]			#5	CommonGIS	decision tasks with geographical components
7	[93]				Data Context Map	unified attribute representation
8	[73]				Decision Exploration Lab	operational decision management
9	[333]			#5	Dissimilarity Prototypes	decisions organised in prototypes
10	[525]			#4	Dust & Magnet	magnet metaphor for multi-attribute data
11	[514]			#2	EZChooser	products as image collection, attributes as bargrams
12	[415]			#1	Financial Portfolio	choose financial portfolio
13	[405]			#1	FinVis	personal financial decisions
14	[443]				FOCUS	interactive table for product comparison
15	[121]				FundExplorer	stock investiment decision support
16	[16, 513]			#3	HomeFinder/FilmFinder	database queries with direct manipulation
17	[311]			#5	KDD-based TDSS	temporal decision support
18	[109]				Lattice graphs	lattice vis in speech recognition to support decisions
19	[201]			#5	LineUp	multi-attribute rankings
20	[439]			#5	LiteVis	choose lighting design configuration
21	[307]			#5	ManyLists	product comparison with spatial layout
22	[497]				NameVoyager	choose baby names
23	[28]			#4	Ocelot	decision support for cyber analysts
24	[233]				Ovis	ocean forecast
25	[142]			#3	PeerFinder	find similar people to guide life choices
26	[207]			#4	PROACT	for prostate cancer health risk communication
27	[396]			#3	Product Explorer	parallel coordinates for product comparison
28	[284]			#5	Run Watchers	flood management decisions (mostly automated)
29	[156]				ScatterDice	multi-attribute data exploration with 3D scatterplots
30	[427]				Shen-Hsieh & Schindl	case study of business strategic decisions
31	[389]			#5	SmartClient	product comparison in electronic commerce
32	[27]				Traffic Origins	aid decisions in traffic management
33	[35]			#4	Stratos	software release plans
34	[161]			#5	Urbane	urban planning decisions
35	[50, 51, 85, 110]			#2	Value Charts	stacked barcharts for preferential choice
36	[518]				Visible Decisions	3D vis capital market decisions
37	[127]			#4	VisIDM	decisions under uncertainty and risk
38	[364]			#5	WeightLifter	weight-based multi-attribute decisions

39	[137]	#1	Parallel Coordinates	chapter 3
40	[137]	#1	Scatterplot Matrix	chapter 3
41	[137]	#1	Tabular Visualization	chapter 3

Bibliography

[1] Phys.org: Study shows banks that have good working relationships with their customers reduce loan defaults.
https://phys.org/news/2014-09-banks-good-relationships-customers-loan.html.
Accessed: 2017-08-2.

[2] Waybackmachine: Clustering illusion.
https://web.archive.org/web/20160320073604/http://bias123.com/clustering_illusion.
Accessed: 2017-08-01.

[3] Colorbrewer 2.0 color advice for cartography.
http://colorbrewer2.org/.
Accessed: 2017-02-17.

[4] Decisive 2017 dealing with cognitive biases in visualisations : a vis 2017 workshop.
http://decisive-workshop.dbvis.de/.
Accessed: 2017-08-01.

[5] MS Windows NT kernel description.
http://www.dictionary.com/browse/paradigm.
Accessed: 2017-08-29.

[6] dutchnews.nl: It is worth asking for a second opinion, patient lobby group says.
http://www.dutchnews.nl/news/archives/2017/02/it-is-worth-asking-for-a-second-opinion-pat
Accessed: 2017-08-2.

[7] Get to know the finder on your mac.
https://support.apple.com/en-us/HT201732.
Accessed: 2017-02-24.

[8] Medium: Director of national intelligence james clapper outlines “litany of doom” in u.s. intelligence community’s annual worldwide threat assessment on capitol hill.
https://medium.com/@ODNIgov/dni-clapper-outlines-litany-of-doom-in-annual-worldwide-threa
.
Accessed: 2017-08-29.

[9] Business insider: These 17 agencies make up the most sophisticated spy network in the world.
http://www.businessinsider.com/17-agencies-of-the-us-intelligence-community-2013-5?IR=T, .
Accessed: 2017-08-2.

[10] Lineup - legatum prosperity index.
https://lineup.caleydoapp.org.
Accessed: 2017-02-24.

[11] ichive.
http://thechive.com/2016/03/04/real-life-optical-illusions-will-make-you-look-more-
Accessed: 2017-08-01.

[12] Ads of the world: Print advertisment created by leo burnett, brazil for ae, within the category: Finance. "don't be lost in the weird world of investments.".
https://adsoftheworld.com/media/print/ae_surreal_2.
Accessed: 2017-08-2.

[13] Treemaps for space-constrained visualization of hierarchies.
http://www.cs.umd.edu/hcil/treemap-history/.
Accessed: 2017-08-16.

[14] Vis conference 2017: Workshop.
http://ieeevis.org/year/2017/info/workshops.
Accessed: 2017-08-01.

[15] Shehzad Afzal, Ross Maciejewski, and David S Ebert.
Visual analytics decision support environment for epidemic modeling and response evaluation.
In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 191–200. IEEE, 2011.

[16] Christopher Ahlberg and Ben Shneiderman.
Visual information seeking: Tight coupling of dynamic query filters with starfield displays.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 313–317. ACM, 1994.

[17] Maurice Allais.
Rational man's behavior in the presence of risk: Critique of the postulates and axioms of the american school.
Econometrica, 21(4):503–46, 1953.

[18] Scott T Allison and David M Messick.
The group attribution error.
Journal of Experimental Social Psychology, 21(6):563–579, 1985.

[19] Robert Amar and John Stasko.
Best paper: A knowledge task-based framework for design and evaluation of information visualizations.
In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 143–150. IEEE, 2004.

[20] Robert Amar, James Eagan, and John Stasko.
Low-level components of analytic activity in information visualization.
In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 111–117. IEEE, 2005.

[21] Robert A Amar and John T Stasko.
Knowledge precepts for design and evaluation of information visualizations.
IEEE Transactions on Visualization and Computer Graphics, 11(4):432–442, 2005.

[22] David F Andrews.
Plots of high-dimensional data.
Biometrics, pages 125–136, 1972.

[23] Gennady Andrienko, Natalia Andrienko, and Piotr Jankowski.
Building spatial decision support tools for individuals and groups.
Journal of Decision Systems, 12(2):193–208, 2003.

[24] Natalia Andrienko and Gennady Andrienko.
Informed spatial decisions through coordinated views.
Information Visualization, 2(4):270–285, 2003.

[25] Nikos Antonopoulos, Andreas Veglis, Antonis Gardikiotis, Rigas Kotsakis, and George Kalliris.
Web third-person effect in structural aspects of the information on media websites.
Computers in Human Behavior, 44:48–58, 2015.

[26] David Antons and Frank T Piller.
Opening the black box of "not invented here": attitudes, decision biases, and behavioral consequences.
The Academy of Management Perspectives, 29(2):193–217, 2015.

[27] Afian Anwar, Till Nagel, and Carlo Ratti.
Traffic origins: A simple visualization technique to support traffic incident analysis.
In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 316–319. IEEE, 2014.

[28] Dustin L Arendt, Russ Burtner, Daniel M Best, Nathan D Bos, John R Gersh, Christine D Piatko, and Celeste Lyn Paul.
Ocelot: user-centered design of a decision support visualization for network quarantine.
In *Visualization for Cyber Security (VizSec), 2015 IEEE Symposium on*, pages 1–8. IEEE, 2015.

[29] Dan Ariely and Thomas S. Wallsten.
Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect.
Organizational Behavior and Human Decision Processes, 63(3):223 – 232, 1995.

[30] Hal R Arkes.
Costs and benefits of judgment errors: Implications for debiasing.
Psychological bulletin, 110(3):486, 1991.

[31] David Arnott.
Cognitive biases and decision support systems development: a design science approach.
Information Systems Journal, 16(1):55–78, 2006.

[32] Kenneth J Arrow.
A difficulty in the concept of social welfare.
Journal of political economy, 58(4):328–346, 1950.

[33] Toshiyuki Asahi, David Turo, and Ben Shneiderman.
Using treemaps to visualize the analytic hierarchy process.
Information Systems Research, 6(4):357–375, 1995.

[34] Solomon E Asch and H Guetzkow.
Effects of group pressure upon the modification and distortion of judgments.
Groups, leadership, and men, pages 222–236, 1951.

[35] Bon Adriel Aseniero, Tiffany Wun, David Ledo, Guenther Ruhe, Anthony Tang, and Sheelagh Carpendale.
Stratos: Using visualization to support decisions in strategic software release planning.
In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1479–1488. ACM, 2015.

[36] Fred Attneave.
Psychological probability as a function of experienced frequency.
Journal of Experimental Psychology, 46(2):81, 1953.

[37] Collective Authors.
Psychology Today: An Introduction (Second Edition).
Comunication Research Machines (CRM), 1972.

[38] Michael S Ayers and Lynne M Reder.
A theoretical review of the misinformation effect: Predictions from an activation-based memory model.
Psychonomic Bulletin & Review, 5(1):1–21, 1998.

[39] Benjamin Bach, Natalie Kerracher, Kyle Wm Hall, Sheelagh Carpendale, Jessie Kennedy, and Nathalie Henry Riche.
Telling stories about dynamic networks with graph comics.
In *Proc. of the Conference on Human Factors in Information Systems (CHI)*. ACM, New York, United States, 2016.

[40] Diane F Baker.
Enhancing group decision making: An exercise to reduce shared information bias.
Journal of Management Education, 34(2):249–279, 2010.

[41] Aron K Barbey and Steven A Sloman.
Base-rate respect: From ecological rationality to dual processes.
Behavioral and Brain Sciences, 30(03):241–254, 2007.

[42] Jonathan Baron and John C Hershey.
Outcome bias in decision evaluation.
Journal of personality and social psychology, 54(4):569, 1988.

[43] Jonathan Baron, Jane Beattie, and John C Hershey.
Heuristics and biases in diagnostic reasoning: Ii. congruence, information, and certainty.
Organizational Behavior and Human Decision Processes, 42(1):88–110, 1988.

[44] Jonathan Baron, Rajeev Gowda, and Howard Kunreuther.
Attitudes toward managing hazardous waste: What should be cleaned up and who should pay for it?
Risk Analysis, 13(2):183–192, 1993.

[45] Lyn Bartram, Abhisekh Patra, and Maureen Stone.
Affective color in visualization.
In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1364–1374. ACM, 2017.

[46] Kaye Enid Basford and John Wilder Tukey.
Graphical Analysis of Multi-Response Data, volume 6.
CRC Press, 1998.

[47] Melissa Bateson, Susan D. Healy, and Hurly.
Irrational choices in hummingbird foraging behaviour.
Animal Behaviour, 63(3):587–596, March 2002.
ISSN 00033472.
doi: 10.1006/anbe.2001.1925.

[48] Jared S Bauer, Sunny Consolvo, Benjamin Greenstein, Jonathan Schooler, Eric Wu, Nathaniel F Watson, and Julie Kientz.
Shuteye: encouraging awareness of healthy sleep recommendations with a mobile, peripheral display.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1401–1410. ACM, 2012.

[49] Dominikus Baur, Frederik Seiffert, Michael Sedlmair, and Sebastian Boring.
The streams of our lives: Visualizing listening histories in context.
IEEE Transactions on Visualization and Computer Graphics, 16(6):1119–1128, 2010.

[50] Jeanette Bautista and Giuseppe Carenini.
An integrated task-based framework for the design and evaluation of visualizations to support preferential choice.
In *Proceedings of the working conference on Advanced visual interfaces*, pages 217–224. ACM, 2006.

[51] Jeanette Bautista and Giuseppe Carenini.
An empirical evaluation of interactive visualizations for preferential choice.
In *Proceedings of the working conference on Advanced visual interfaces*, pages 207–214. ACM, 2008.

[52] Richard A Becker and William S Cleveland.
Brushing scatterplots.
Technometrics, 29(2):127–142, 1987.

[53] Richard A. Becker, Stephen G. Eick, and Allan R. Wilks.
Visualizing network data.
IEEE Transactions on visualization and computer graphics, 1(1):16–28, 1995.

[54] David E Bell.
Disappointment in decision making under uncertainty.
Operations research, 33(1):1–27, 1985.

[55] Palash Bera.
How colors in business dashboards affect users' decision making.
Communications of the ACM, 59(4):50–57, 2016.

[56] Joseph Berkson.
Limitations of the application of fourfold table analysis to hospital data.
Biometrics Bulletin, 2(3):47–53, 1946.

[57] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich.
Soylent: a word processor with a crowd inside.
Communications of the ACM, 58(8):85–94, 2015.

[58] Jacques Bertin.
La graphique et le traitement graphique de l'information.
Number 91 (084.21) BER. 1977.

[59] James R. Bettman, Mary F. Luce, and John W. Payne.
Constructive consumer choice processes.
J. of Consumer Research, 25(3):187–217, December 1998.

[60] William M Bolstad.
Introduction to Bayesian statistics.
J. Wiley & Sons, 2013.

[61] Sarah E Bonner, Reid Hastie, Geoffrey B Sprinkle, and S Mark Young.
A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting.
Journal of Management Accounting Research, 12(1):19–64, 2000.

[62] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva.
Beyond memorability: Visualization recognition and recall.
IEEE transactions on visualization and computer graphics, 22(1):519–528, 2016.

[63] Nadia Boukhelifa, Anastasia Bezerianos, Tobias Isenberg, and Jean-Daniel Fekete.
Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty.
TVCG, 18(12):2769–2778, 2012.

[64] Liliana Bounegru, Tommaso Venturini, Jonathan Gray, and Mathieu Jacomy.
Narrating networks: Exploring the affordances of networks as storytelling devices in journalism.
Digital Journalism, pages 1–32, 2016.

[65] J. Boy, R.A. Rensink, E. Bertini, and J.-D. Fekete.
A principled way of assessing visualization literacy.
Visualization and Computer Graphics, IEEE Transactions on, 20(12):1963–1972, Dec 2014.
ISSN 1077-2626.
doi: 10.1109/TVCG.2014.2346984.

[66] Jeremy Boy, Francoise Detienne, and Jean-Daniel Fekete.
Storytelling in information visualizations: Does it engage users to explore data?
In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1449–1458. ACM, 2015.

[67] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini.
Showing people behind data: Does anthropomorphizing visualizations elicit more empathy for human rights data?
In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5462–5474. ACM, 2017.

[68] Charles J Brainerd and Valerie F Reyna.
Fuzzy-trace theory and false memory.
Current Directions in Psychological Science, 11(5):164–169, 2002.

[69] Sharon S Brehm.
Psychological reactance and the attractiveness of unobtainable objects: Sex differences in children's responses to an elimination of freedom.
Sex Roles, 7(9):937–949, 1981.

[70] M. Brehmer and T. Munzner.
A multi-level typology of abstract visualization tasks.
IEEE Transactions on Visualization and Computer Graphics, 19(12):2376–2385, Dec 2013.
ISSN 1077-2626.
doi: 10.1109/TVCG.2013.124.

[71] Matthew Brehmer, Jocelyn Ng, Kevin Tate, and Tamara Munzner.
Matches, mismatches, and methods: multiple-view workflows for energy portfolio analysis.
IEEE transactions on visualization and computer graphics, 22(1):449–458, 2016.

[72] Malcolm Brenner.
The next-in-line effect.
Journal of Verbal Learning and Verbal Behavior, 12(3):320–323, 1973.

[73] Bertjan Broeksema, Thomas Baudel, Alex Telea, and Paolo Crisafulli.
Decision exploration lab: A visual analytics solution for decision management.
IEEE Transactions on Visualization and Computer Graphics, 19(12):1972–1981, 2013.

[74] Alan S Brown.
A review of the tip-of-the-tongue experience.
Psychological bulletin, 109(2):204, 1991.

[75] Alan S Brown and Dana R Murphy.
Cryptomnesia: Delineating inadvertent plagiarism.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(3):432, 1989.

[76] Christina L Brown.
"do the right thing:" diverging effects of accountability in a managerial context.
Marketing Science, 18(3):230–246, 1999.

[77] Jonas Buddeberg, Bernhard Jenny, and Wesley Willett.
Interactive shearing for terrain visualization: an expert study.
GeoInformatica, 21(3):643–665, 2017.

[78] Roger Buehler, Dale Griffin, and Johanna Peetz.
Chapter one-the planning fallacy: cognitive, motivational, and social origins.
Advances in experimental social psychology, 43:1–62, 2010.

[79] Maria Rosaria Cadinu and Myron Rothbart.
Self-anchoring and differentiation processes in the minimal group setting.
Journal of personality and social psychology, 70(4):661, 1996.

[80] Colin Camerer.
Behavioral game theory: Experiments in strategic interaction.
Princeton University Press, 2003.

[81] Colin Camerer, George Loewenstein, and Martin Weber.
The curse of knowledge in economic settings: An experimental analysis.
Journal of political Economy, 97(5):1232–1254, 1989.

[82] Colin F Camerer, Robin M Hogarth, David V Budescu, and Catherine Eckel.
The effects of financial incentives in experiments: A review and capital-labor-production framework.
In *Elicitation of Preferences*, pages 7–48. Springer, 1999.

[83] W Keith Campbell and Constantine Sedikides.
Self-threat magnifies the self-serving bias.
Review of General Psychology, 3(1):23–43, 1999.

[84] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors.
Readings in Information Visualization: Using Vision to Think.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.
ISBN 1-55860-533-9.

[85] Giuseppe Carenini and John Loyd.
Valuecharts: analyzing linear models expressing preferences and evaluations.
In *Proceedings of the working conference on Advanced visual interfaces*, pages 150–157. ACM, 2004.

[86] Megan A Carney.
"sharing one's destiny": Effects of austerity on migrant health provisioning in the mediterranean borderlands.
Social Science & Medicine, 2017.

[87] Daniel B Carr, Richard J Littlefield, WL Nicholson, and JS Littlefield.
Scatterplot matrix techniques for large n.
Journal of the American Statistical Association, 82(398):424–436, 1987.

[88] Melanie Cary and Lynne M Reder.
A dual-process account of the list-length and strength-based mirror effects in recognition.
Journal of Memory and Language, 49(2):231–248, 2003.

[89] Loren J Chapman and Jean P Chapman.
Illusory correlation as an obstacle to the use of valid psychodiagnostic signs.
Journal of abnormal psychology, 74(3):271, 1969.

[90] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos.
Apolo: making sense of large network data by combining rich user interaction and machine learning.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 167–176. ACM, 2011.

[91] Y. Chen, S. Barlowe, and J. Yang.
Click2annotate: Automated insight externalization with rich semantics.
In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 155–162, Oct 2010.
doi: 10.1109/VAST.2010.5652885.

[92] PN Cheng, KJ Holyoak, and RE Nisbett.
Pragmatic versus syntactic approaches to training deductive reasoning.
Cognitive Psychology, 18:293–328, 1986.

[93] Shenghui Cheng and Klaus Mueller.
The data context map: Fusing data and attributes into a unified display.
Visualization and Computer Graphics, IEEE Transactions on, 22(1):121–130, 2016.

[94] Alexander Chernev.
When more is less and less is more: The role of ideal point availability and assortment in consumer choice.
Journal of Consumer Research, 30(2):170–183, 2003.

[95] Herman Chernoff.
The use of faces to represent points in k-dimensional space graphically.
Journal of the American Statistical Association, 68(342):361–368, 1973.

[96] Herman Chernoff and M Haseeb Rizvi.
Effect on classification error of random permutations of features in representing multivariate data by faces.
Journal of the American Statistical Association, 70(351a):548–554, 1975.

[97] Parmit K Chilana, Jacob O Wobbrock, and Andrew J Ko.
Understanding usability practices in complex domains.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2337–2346. ACM, 2010.

[98] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou.
The anchoring effect in decision-making with visual analytics.
IEEE Transactions on Visualization and Computer Graphics, 2018.

[99] Jongwoon Willie Choi, Gary W Hecht, and William B Tayler.
Strategy selection, surrogation, and strategic performance measurement systems.
Journal of Accounting Research, 51(1):105–133, 2013.

[100] H Choo and SL Franconeri.
Enumeration of small collections violates weber's law.
Psychonomic bulletin & review, 21(1):93–99, 2014.

[101] Jessica M. Choplin and John E. Hummel.
Comparison-induced decoy effects.
Memory & Cognition, 33(2):332–343, 2005.
ISSN 0090-502X.

[102] Mei C Chuah and Steven F Roth.
On the semantics of interactive visualizations.
In *Information Visualization'96, Proceedings IEEE Symposium on*, pages 29–36. IEEE, 1996.

[103] Benjamin A Clegg, Rosa Mikeal Martey, Jennifer Stromer-Galley, Kate Kenski, Tobi Saulnier, James E Folkestad, Elizabeth McLaren, Adrienne Shaw, Joanna E Lewis, John D Patterson, et al.
Game-based training to mitigate three forms of cognitive bias.
In *Proceedings of Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, volume 14180, pages 1–12, 2014.

[104] William S Cleveland.
Visualizing data.
Hobart Press, 1993.

[105] William S Cleveland and Robert McGill.
Graphical perception: Theory, experimentation, and application to the development of graphical methods.
Journal of the American statistical association, 79(387):531–554, 1984.

[106] Jacob Cohen.
The Earth is round (p<.05).
American psychologist, 49(12):997, 1994.

[107] John D Coley, Melanie Arenson, Yian Xu, and Kimberly D Tanner.
Intuitive biological thought: Developmental changes and effects of biology education in late adolescence.
Cognitive psychology, 92:1–21, 2017.

[108] Open Science Collaboration et al.
Estimating the reproducibility of psychological science.
Science, 349(6251), 2015.
doi: 10.1126/science.aac4716.

[109] Christopher Collins, M Sheelagh T Carpendale, and Gerald Penn.
Visualization of uncertainty in lattices to support decision-making.
In *EuroVis*, pages 51–58, 2007.

[110] Cristina Conati, Giuseppe Carenini, Enamul Hoque, Ben Steichen, and Dereck Toker.
Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making.
In *Computer Graphics Forum*, volume 33, pages 371–380. Wiley Online Library, 2014.

[111] Mariano Consens and Alberto Mendelzon.
Hy+: A hygraph-based query and visualization system.
In *ACM SIGMOD Record*, volume 22, pages 511–516. ACM, 1993.

[112] CH Coombs.
Portfolio theory and the measurement of risk. en mf kaplan & s. schwartz (eds). human judgment and deci-sion processes.
1975.

[113] Joshua Correll, Bernadette Park, Charles M Judd, and Bernd Wittenbrink.
The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals.
Journal of personality and social psychology, 83(6):1314, 2002.

[114] Michael Correll and Jeffrey Heer.
Regression by eye: Estimating trends in bivariate visualizations.
In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1387–1396. ACM, 2017.

[115] Fergus IM Craik and Endel Tulving.
Depth of processing and the retention of words in episodic memory.
Journal of experimental Psychology: general, 104(3):268, 1975.

[116] Paolo Crosetto and Alexia Gaudeul.
Testing the strength and robustness of the attraction effect in consumer decision making.
2015.

[117] Pat Croskerry, Geeta Singhal, and Sílvia Mamede.
Cognitive debiasing 1: origins of bias and theory of debiasing.
BMJ quality & safety, 22(Suppl 2):ii58–ii64, 2013.

[118] Colin Crouch.
Post-democracy.
Polity Cambridge, 2004.

[119] Douglas P Crowne and David Marlowe.
A new scale of social desirability independent of psychopathology.
Journal of consulting psychology, 24(4):349, 1960.

[120] Robert J Crutcher and Alice F Healy.
Cognitive operations and the generation effect.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4):669, 1989.

[121] Christoph Csallner, Marcus Handte, Othmar Lehmann, and John Stasko.
Fundexplorer: Supporting the diversification of mutual fund portfolios using context treemaps.
In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 203–208. IEEE, 2003.

[122] Pepper D Culpepper.
The political economy of unmediated democracy: Italian austerity under mario monti.
West European Politics, 37(6):1264–1281, 2014.

[123] Geoff Cumming and Sue Finch.
Inference by eye: confidence intervals and how to read pictures of data.
American Psychologist, 60(2):170, 2005.

[124] Mary Cummings.
Automation bias in intelligent time critical decision support systems.
In *AIAA 1st Intelligent Systems Technical Conference*, page 6313, 2004.

[125] Fiery Cushman.
Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment.
Cognition, 108(2):353–380, 2008.

[126] Tuan Nhon Dang, Anushka Anand, and Leland Wilkinson.
Timeseer: Scagnostics for high-dimensional time series.
IEEE Transactions on Visualization and Computer Graphics, 19(3):470–483, 2013.

[127] Mohammad Daradkeh, Clare Churcher, and Alan McKinnon.
Supporting informed decision-making under uncertainty and risk through interactive visualisation.
In *Proceedings of the Fourteenth Australasian User Interface Conference-Volume 139*, pages 23–32. Australian Computer Society, Inc., 2013.

[128] Aritra Dasgupta, Jorge Poco, Yaxing Wei, Robert Cook, Enrico Bertini, and Claudio T Silva.
Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison.
IEEE transactions on visualization and computer graphics, 21(9):996–1014, 2015.

[129] Robyn M Dawes, David Faust, and Paul E Meehl.
Clinical versus actuarial judgment.
Science, 243(4899):1668–1674, 1989.

[130] Antoine de Falguerolles, Felix Friedrich, and Günther Sawitzki.
A tribute to J. Bertin's graphical data analysis.
In W. Bandilla and F. Faulbaum, editors, *SoftStat '97 (Advances in Statistical Software 6)*, pages 11–20. Lucius & Lucius, 1997.

[131] Ravi Dhar and Itamar Simonson.
The effect of forced choice on choice.
Journal of Marketing Research, 40(2):146–160, 2003.

[132] Evanthia Dimara, Pierre Dragicevic, and Anastasia Bezerianos.
Accounting for availability biases in information visualization.
In *IEEE VIS 2014*, 2014.

[133] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic.
Testing the attraction effect on two information visualization datasets.
Research Report RR-8895, Inria, March 2016.
URL `http://hal.inria.fr/hal-01295624`.

[134] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic.
The attraction effect in information visualization.
IEEE Transactions on Visualization and Computer Graphics, 23(1):471–480, 2017.

[135] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic.
Narratives in crowdsourced evaluation of visualizations: A double-edged sword?
In *ACM Conference on Human Factors in Computing Systems (CHI)*, Denver, United States, May 2017.
doi: 10.1145/3025453.3025870.

[136] Evanthia Dimara, Paola Valdivia, and Christoph Kinkeldey.
Dcpairs: A pairs plot based decision support system.
In *EuroVis-19th EG/VGTC Conference on Visualization*, 2017.

[137] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic.
Conceptual and methodological issues in evaluating multidimensional visualizations for decision support.
IEEE Transactions on Visualization and Computer Graphics, X(X):XXX–XXX, 2018.

[138] Kenneth A Dodge.
Translational science in action: Hostile attributional style and the development of aggressive behavior problems.
Development and psychopathology, 18(03):791–814, 2006.

[139] Julie S Downs.
Prescriptive scientific narratives for communicating usable science.
Proceedings of the National Academy of Sciences, 111(Supplement 4):13627–13633, 2014.

[140] Pierre Dragicevic.
Fair statistical communication in hci.
In *Modern Statistical Methods for HCI*, pages 291–330. Springer, 2016.

[141] Pierre Dragicevic and Yvonne Jansen.
Visualization-mediated alleviation of the planning fallacy.
In *IEEE VIS 2014*, 2014.

[142] Fan Du, Catherine Plaisant, Neil Spring, and Ben Shneiderman.
Finding similar people to guide life choices: Challenge, design, and evaluation.
In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5498–5544. ACM, 2017.

[143] Norah E Dunbar, Matthew L Jensen, Claude H Miller, Elena Bessarabova, Sara K Straub, Scott N Wilson, Javier Elizondo, Judee K Burgoon, Joseph S Valacich, Bradley Adame, et al.
Mitigating cognitive bias through the use of serious games: Effects of feedback.
In *International Conference on Persuasive Technology*, pages 92–105. Springer, 2014.

[144] Luce R Duncan.
Individual choice behavior.
A theoretical analysis, New York, London, 1959.

[145] David Dunning, Judith A Meyerowitz, and Amy D Holzberg.
Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability.
Journal of personality and social psychology, 57(6):1082, 1989.

[146] Gilles Dutilh and Jörg Rieskamp.
Comparing perceptual and preferential decision making.
Psychonomic bulletin & review, 23(3):723–737, 2016.

[147] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang.
Explicit warnings reduce but do not eliminate the continued influence of misinformation.
Memory & cognition, 38(8):1087–1100, 2010.

[148] Charalampos Economou, Daphne Kaitelidou, Dimitris Katsikas, Olga Siskou, and Maria Zafiropoulou.
Impacts of the economic crisis on access to healthcare services in greece with a focus on the vulnerable groups of the population.
Social Cohesion and Development, 9(2), 2014.

[149] Ward Edwards.
Conservatism in human information processing.
Formal representation of human judgment, 17:51, 1968.

[150] Serge Egelman, Ed H Chi, and Steven Dow.
Crowdsourcing in hci research.
In *Ways of Knowing in HCI*, pages 267–289. Springer, 2014.

[151] Hillel J Einhorn and Robin M Hogarth.
Behavioral decision theory: Processes of judgement and choice.
Annual review of psychology, 32(1):53–88, 1981.

[152] Gilles O Einstein, Mark A McDaniel, Carrie L Williford, Jason L Pagan, and R Dismukes.
Forgetting of intentions in demanding situations is rapid.
Journal of Experimental Psychology: Applied, 9(3):147, 2003.

[153] Micheline Elias and Anastasia Bezerianos.
Annotating bi visualization dashboards: Needs & challenges.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1641–1650. ACM, 2012.

[154] Geoffrey Ellis and Alan Dix.
Decision making under uncertainty in visualisation?
In *IEEE VIS2015*, 2015.

[155] Daniel Ellsberg.
Risk, ambiguity, and the savage axioms.
The quarterly journal of economics, pages 643–669, 1961.

[156] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete.
Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation.
IEEE transactions on Visualization and Computer Graphics, 14(6):1539–1148, 2008.

[157] John W Emerson, Walton A Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, and Hadley Wickham.
The generalized pairs plot.
Journal of Computational and Graphical Statistics, 22(1):79–91, 2013.

[158] J St BT Evans, Julie L Barston, and Paul Pollard.
On the conflict between logic and belief in syllogistic reasoning.
Memory & cognition, 11(3):295–306, 1983.

[159] J-D Fekete.
The infovis toolkit.
In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 167–174. IEEE, 2004.

[160] Jean-Daniel Fekete, Jarke J Van Wijk, John T Stasko, and Chris North.
The value of information visualization.
In *Information visualization*, pages 1–18. Springer, 2008.

[161] Nivan Ferreira, Marcos Lage, Harish Doraiswamy, Huy Vo, Luc Wilson, Heidi Werner, Muchan Park, and Cláudio Silva.
Urbane: A 3d framework to support data driven decision making in urban development.
In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pages 97–104. IEEE, 2015.

[162] Stephen Few.
Show me the numbers: Designing tables and graphs to enlighten.
Analytics Press, 2012.

[163] Fabian Fischer and Daniel Keim.
Vacs: Visual analytics suite for cyber security-visual exploration of cyber security datasets.
In *IEEE VIS*, 2013.

[164] B. Fischhoff.
Debiasing/kahneman, d., slovic, p. and tversky, a.
In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge University Press, 1982.

[165] Baruch Fischhoff and Ruth Beyth.
I knew it would happen: Remembered probabilities of once—future things.
Organizational Behavior and Human Performance, 13(1):1–16, 1975.

[166] Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein.
Knowing with certainty: The appropriateness of extreme confidence.
Journal of Experimental Psychology: Human perception and performance, 3(4):552, 1977.

[167] Baruch Fischhoff, Cherie Chauvin, and Editors.
Intelligence analysis: Behavioral and social scientific foundations.
National Academies Press, 2011.

[168] Peter C Fishburn.
Ssb utility theory and decision-making under uncertainty.
Mathematical social sciences, 8(3):253–285, 1984.

[169] Robert J Fisher.
Social desirability bias and the validity of indirect questioning.
Journal of consumer research, 20(2):303–315, 1993.

[170] Susan T Fiske.
Attention and weight in person perception: The impact of negative and extreme behavior.
Journal of personality and Social Psychology, 38(6):889, 1980.

[171] Geoffrey T Fong and Richard E Nisbett.
Immediate and delayed transfer of training effects in statistical reasoning.
Journal of Experimental Psychology: General, 120(1):34, 1991.

[172] Geoffrey T Fong, David H Krantz, and Richard E Nisbett.
The effects of statistical training on thinking about everyday problems.
Cognitive psychology, 18(3):253–292, 1986.

[173] Robert H Frank.
Passions within reason: the strategic role of the emotions.
WW Norton & Co, 1988.

[174] Shane Frederick, Leonard Lee, and Ernest Baskin.
The limits of attraction.
Journal of Marketing Research, 51(4):487–507, 2014.
doi: 10.1509/jmr.12.0061.

[175] Barbara L Fredrickson and Daniel Kahneman.
Duration neglect in retrospective evaluations of affective episodes.
Journal of personality and social psychology, 65(1):45, 1993.

[176] Deborah Frisch and Jonathan Baron.
Ambiguity and rationality.
Journal of Behavioral Decision Making, 1(3):149–157, 1988.

[177] Jon Froehlich, Leah Findlater, Marilyn Ostergren, Solai Ramanathan, Josh Peterson, Inness Wragg, Eric Larson, Fabia Fu, Mazhengmin Bai, Shwetak Patel, et al.
The design and evaluation of prototype eco-feedback displays for fixture-level water usage data.
In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2367–2376. ACM, 2012.

[178] Ying-Huey Fua, Matthew O Ward, and Elke A Rundensteiner.
Hierarchical parallel coordinates for exploration of large datasets.
In *Proceedings of the conference on Visualization'99: celebrating ten years*, pages 43–50. IEEE Computer Society Press, 1999.

[179] Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, and Daniel Keim.
A systematic review of experimental studies on data glyphs.
IEEE transactions on visualization and computer graphics, 2016.

[180] George W Furnas and Andreas Buja.
Prosection views: Dimensional inference through sections and projections.
Journal of Computational and Graphical Statistics, 3(4):323–353, 1994.

[181] Adrian Furnham and Hua Chu Boo.
A literature review of the anchoring effect.
The Journal of Socio-Economics, 40(1):35–42, 2011.

[182] Adrian Furnham and Sandra Schofield.
Accepting personality test feedback: A review of the barnum effect.
Current Psychology, 6(2):162–178, 1987.

[183] Stephen M Garcia, Hyunjin Song, and Abraham Tesser.
Tainted recommendations: The social comparison bias.
Organizational Behavior and Human Decision Processes, 113(2):97–101, 2010.

[184] Andrew B Geier, Paul Rozin, and Gheorghe Doros.
Unit bias: A new heuristic that helps explain the effect of portion size on food intake.
Psychological Science, 17(6):521–525, 2006.

[185] Tim P German and H Clark Barrett.
Functional fixedness in a technologically sparse culture.
Psychological Science, 16(1):1–5, 2005.

[186] John R Gersh, Jennifer A McKneely, and Roger W Remington.
Cognitive engineering: Understanding human interaction with complex systems.
Johns Hopkins APL technical digest, 26(4):377–382, 2005.

[187] Jeffrey A Gibbons, Andrew K Velkey, and Kathren T Partin.
Influence of recall procedures on the modality effect with numbers and enumerated stimuli.
The Journal of general psychology, 135(1):84–104, 2008.

[188] Brian J Gibbs and Aimee Drolet.
Consumption effort: The mental cost of generating utility and the role of consumer energy level in ambitious consumption.
Journal of Consumer Psychology, 13(3):268–277, 2003.

[189] Gerd Gigerenzer.
How to make cognitive illusions disappear: Beyond "heuristics and biases".
European review of social psychology, 2(1):83–115, 1991.

[190] Gerd Gigerenzer.
Why heuristics work.
Perspectives on psychological science, 3(1):20–29, 2008.

[191] Gerd Gigerenzer.
Risk savvy: How to make good decisions.
Penguin, 2015.

[192] Gerd Gigerenzer and Henry Brighton.
Homo heuristicus: Why biased minds make better inferences.
Topics in cognitive science, 1(1):107–143, 2009.

[193] Daniel T Gilbert and Patrick S Malone.
The correspondence bias.
Psychological bulletin, 117(1):21, 1995.

[194] Daniel T Gilbert, Ryan P Brown, Elizabeth C Pinel, and Timothy D Wilson.
The illusion of external agency.
Journal of personality and social psychology, 79(5):690, 2000.

[195] Thomas Gilovich, Robert Vallone, and Amos Tversky.
The hot hand in basketball: On the misperception of random sequences.
Cognitive psychology, 17(3):295–314, 1985.

[196] Thomas Gilovich, Victoria Husted Medvec, and Kenneth Savitsky.
The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance.
Journal of personality and social psychology, 78(2):211, 2000.

[197] Paul Ginns.
Meta-analysis of the modality effect.
Learning and Instruction, 15(4):313–331, 2005.

[198] Nada Golmie and Bill Kules.
Highlight and selection control for dynamic table visualization.
CMSC 838S Term Paper, 1999.

[199] Mark L Graber, Stephanie Kissam, Velma L Payne, Ashley ND Meyer, Asta Sorensen, Nancy Lenfestey, Elizabeth Tant, Kerm Henriksen, Kenneth LaBresh, and Hardeep Singh.
Cognitive interventions to reduce diagnostic error: a narrative review.
BMJ Quality & Safety, pages bmjqs–2011, 2012.

[200] Sandra Graham, Cynthia Hudley, and Estella Williams.
Attributional and emotional determinants of aggression among african-american and latino young adolescents.
Developmental Psychology, 28(4):731, 1992.

[201] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit.
Lineup: Visual analysis of multi-attribute rankings.
IEEE transactions on visualization and computer graphics, 19(12):2277–2286, 2013.

[202] Melanie C Green and Timothy C Brock.
The role of transportation in the persuasiveness of public narratives.
JPSP, 79(5):701, 2000.

[203] Melanie C Green, Jeffrey J Strange, and Timothy C Brock.
Narrative impact: Social and cognitive foundations.
Taylor & Francis, 2003.

[204] Robert L Greene.
Spacing effects in memory: Evidence for a two-process account.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(3):371, 1989.

[205] Jürgen Habermas.
The lure of technocracy.
2015.

[206] Carolyn L Hafer and Laurent Bègue.
Experimental research on just-world theory: problems, developments, and future challenges.
Psychological bulletin, 131(1):128, 2005.

[207] Anzu Hakone, Lane Harrison, Alvitta Ottley, Nathan Winters, Caitlin Gutheil, Paul KJ Han, and Remco Chang.
Proact: Iterative design of a patient-centered visualization for effective prostate cancer health risk communication.
IEEE transactions on visualization and computer graphics, 23(1):601–610, 2017.

[208] Ruth Hamill, Timothy D Wilson, and Richard E Nisbett.
Insensitivity to sample bias: Generalizing from atypical cases.
Journal of Personality and Social Psychology, 39(4):578, 1980.

[209] Andrew S. Hanks, David Just, and Brian Wansink.
Trigger foods: The influence of "irrelevant" alternatives in school lunchrooms.
Working Papers 05, Cornell Center for Behavioral Economics in Child Nutrition Programs, 2012.

[210] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang.
Ranking visualizations of correlation using weber's law.
IEEE transactions on visualization and computer graphics, 20(12):1943–1952, 2014.

[211] John A Hartigan and Beat Kleiner.
A mosaic of television ratings.
The American Statistician, 38(1):32–35, 1984.

[212] Martie G Haselton.
The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events.
Journal of Research in Personality, 37(1):34–47, 2003.

[213] Lynn Hasher, David Goldstein, and Thomas Toppino.
Frequency and the conference of referential validity.
Journal of verbal learning and verbal behavior, 16(1):107–112, 1977.

[214] Albert H Hastorf and Hadley Cantril.
They saw a game; a case study.
The Journal of Abnormal and Social Psychology, 49(1):129, 1954.

[215] Christopher Healey and James Enns.
Attention and visual memory in visualization and computer graphics.
IEEE transactions on visualization and computer graphics, 18(7):1170–1188, 2012.

[216] Chip Heath.
On the social psychology of agency relationships: Lay theories of motivation overemphasize extrinsic incentives.
Organizational behavior and human decision processes, 78(1):25–62, 1999.

[217] Timothy B. Heath and Subimal Chatterjee.
Asymmetric decoy effects on lower-quality versus higher-quality brands: Meta-analytic and experimental evidence.
Journal of Consumer Research, 22(3):268–284, 1995.

[218] William Hedgcock and Akshay R Rao.
Trade-off aversion as an explanation for the attraction effect: A functional magnetic resonance imaging study.
Journal of Marketing Research, 46(1):1–13, 2009.

[219] William Hedgcock, Akshay R. Rao, and Haipeng (Allan) Chen.
Could ralph nader's entrance and exit have helped al gore? the impact of decoy dynamics on consumer choice.
Journal of Marketing Research, 46(3):330–343, 2009.

[220] James Hedlund.
Risky business: safety regulations, risk compensation, and individual behavior.
Injury prevention, 6(2):82–89, 2000.

[221] Jeffrey Heer and Michael Bostock.
Crowdsourcing graphical perception: using mechanical turk to assess visualization design.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212. ACM, 2010.

[222] Jeffrey Heer and Danah Boyd.
Vizster: Visualizing online social networks.
In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39. IEEE, 2005.

[223] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg.
Voyagers and voyeurs: supporting asynchronous collaborative information visualization.
In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 1029–1038. ACM, 2007.

[224] Julian Heinrich, Yuan Luo, Arthur E. Kirkpatrick, and Daniel Weiskopf.
Evaluation of a bundling technique for parallel coordinates.
Proceedings of International Conference on Information Visualization Theory and Applications, pages 594–602, 2012.

[225] Linda A Henkel and Kimberly J Coffman.
Memory distortions in coerced false confessions: A source monitoring framework analysis.
Applied Cognitive Psychology, 18(5):567–588, 2004.

[226] Nathalie Henry and Jean-Daniel Fekete.
Matrixexplorer: a dual-representation system to explore social networks.
IEEE transactions on visualization and computer graphics, 12(5):677–684, 2006.

[227] Ivan Hernandez and Jesse Lee Preston.
Disfluency disrupts the confirmation bias.
Journal of Experimental Social Psychology, 49(1):178–182, 2013.

[228] Kaisa Herne.
Decoy alternatives in policy choices: Asymmetric domination and compromise effects.
European Journal of Political Economy, 13(3):575 – 589, 1997.
Tiberias Papers.

[229] Kaisa Herne.
The effects of decoy gambles on individual choice.
Experimental Economics, 2(1):31–40, 1999.

[230] Scott Highhouse.
Context-dependent selection: The effects of decoy and phantom job candidates.
Organizational Behavior and Human Decision Processes, 65(1):68 – 76, 1996.
ISSN 0749-5978.

[231] Patrick E Hoffman.
Table visualizations: a formal model and its applications.
PhD thesis, University of Massachusetts Lowell, 1977.

[232] Robin M Hogarth.
A note on aggregating opinions.
Organizational behavior and human performance, 21(1):40–46, 1978.

[233] Thomas Hollt, Ahmed Magdy, Peng Zhan, Guoning Chen, Ganesh Gopalakrishnan, Ibrahim Hoteit, Charles D Hansen, and Markus Hadwiger.
Ovis: A framework for visual analysis of ocean forecast ensembles.
Visualization and Computer Graphics, IEEE Transactions on, 20(8):1114–1126, 2014.

[234] Danny Holten and Jarke J Van Wijk.
Evaluation of cluster identification performance for different pcp variants.
In *Computer Graphics Forum*, volume 29, pages 793–802. Wiley Online Library, 2010.

[235] John Joseph Horton and Lydia B Chilton.
The labor economics of paid crowdsourcing.
In *Proc. of the 11th ACM conference on Electronic commerce*, pages 209–218. ACM, 2010.

[236] Christopher K Hsee and Jiao Zhang.
Distinction bias: misprediction and mischoice due to joint evaluation.
Journal of personality and social psychology, 86(5):680, 2004.

[237] CHRISTOPHER K HSEEÃ.
Less is better: When low-value options are valued more highly than high-value options.
Journal of Behavioral Decision Making, 11:107–121, 1998.

[238] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury.
Personal visualization and personal visual analytics.
IEEE Transactions on Visualization and Computer Graphics, 21(3):420–433, 2015.

[239] Joel Huber and Christopher Puto.
Market boundaries and product choice: Illustrating attraction and substitution effects.
Journal of Consumer Research, pages 31–44, 1983.

[240] Joel Huber, John W Payne, and Christopher Puto.
Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis.
Journal of Consumer Research, 9(1):90–98, 1982.

[241] Joel Huber, John W. Payne, and Christopher P. Puto.
Let's be honest about the attraction effect.
Journal of Marketing Research, 51(4):520–525, 2014.
doi: 10.1509/jmr.14.0208.

[242] Jessica Hullman and Nick Diakopoulos.
Visualization rhetoric: Framing effects in narrative visualization.
TVCG, 17(12):2231–2240, 2011.

[243] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar.
Contextifier: Automatic generation of annotated stock visualizations.
In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2707–2716. ACM, 2013.

[244] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar.
A deeper understanding of sequence in narrative visualization.
TVCG, 19(12):2406–2415, 2013.

[245] Jessica R Hullman.
Not all hits are created equal: Controlling for reasoning and learning processes in mturk.
In *Proceedings of the ACM CHI Workshop on Crowdsourcing and Human Computation*. Citeseer, 2011.

[246] Jean-François Im, Michael J McGuffin, and Rock Leung.
Gplom: the generalized plot matrix for visualizing multidimensional multivariate data.
IEEE Transactions on Visualization and Computer Graphics, 19(12):2606–2614, 2013.

[247] Stephen Ingram, Tamara Munzner, and Marc Olano.
Glimmer: Multilevel mds on the gpu.
IEEE Transactions on Visualization and Computer Graphics, 15(2):249–261, 2009.

[248] Alfred Inselberg.
The plane with parallel coordinates.
The visual computer, 1(2):69–91, 1985.

[249] Petra Isenberg, Anthony Tang, and Sheelagh Carpendale.
An exploratory study of visual information analysis.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1217–1226, New York, NY, USA, 2008. ACM.
ISBN 978-1-60558-011-1.
doi: 10.1145/1357054.1357245.

[250] W Burke Jackson and James V Jucker.
An empirical study of travel time variability and travel choice behavior.
Transportation Science, 16(4):460–475, 1982.

[251] Amy C Janes, Diego A Pizzagalli, Sarah Richardt, Blaise de B Frederick, Avram J Holmes, Jessica Sousa, Maurizio Fava, A Eden Evins, and Marc J Kaufman.
Neural substrates of attentional bias for smoking-related cues: an fmri study.
Neuropsychopharmacology, 35(12):2339–2345, 2010.

[252] Ashok Jansari and Alan J Parkin.
Things that go bump in your life: Explaining the reminiscence bump in autobiographical memory.
Psychology and Aging, 11(1):85, 1996.

[253] Steve MJ Janssen, Antonio G Chessa, and Jaap MJ Murre.
Memory for time: How people date events.
Memory & cognition, 34(1):138–147, 2006.

[254] Jon Jecker and David Landy.
Liking a person as a function of doing him a favour.
Human Relations, 22(4):371–378, 1969.

[255] Karen Jenni and George Loewenstein.
Explaining the identifiable victim effect.
Journal of Risk and Uncertainty, 14(3):235–257, 1997.

[256] Antonio Jiménez, Sixto Rıos-Insua, and Alfonso Mateos.
A decision support system for multiattribute utility evaluation based on imprecise assignments.
Decision Support Systems, 36(1):65–79, 2003.

[257] Jimmy Johansson and Camilla Forsell.
Evaluation of parallel coordinates: Overview, categorization and guidelines for future research.
IEEE transactions on visualization and computer graphics, 22(1):579–588, 2016.

[258] Jimmy Johansson, Patric Ljung, Mikael Jern, and Matthew Cooper.
Revealing structure in visualizations of dense 2d and 3d parallel coordinates.
Information Visualization, 5(2):125–136, 2006.

[259] Hollyn M Johnson and Colleen M Seifert.
Sources of the continued influence effect: When misinformation in memory affects later inferences.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(6):1420, 1994.

[260] Edward Ellsworth Jones and Richard E Nisbett.
The actor and the observer: Divergent perceptions of the causes of behavior.
1971.

[261] John T Jost, Brett W Pelham, and Mauricio R Carvallo.
Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups.
Journal of Experimental Social Psychology, 38(6):586–602, 2002.

[262] Daniel Kahneman.
A perspective on judgment and choice: mapping bounded rationality.
American psychologist, 58(9):697, 2003.

[263] Daniel Kahneman.
Thinking, fast and slow.
Macmillan, 2011.

[264] Daniel Kahneman and Amos Tversky.
Subjective probability: A judgment of representativeness.
Cognitive psychology, 3(3):430–454, 1972.

[265] Daniel Kahneman and Amos Tversky.
Prospect theory: An analysis of decision under risk.
Econometrica: Journal of the econometric society, pages 263–291, 1979.

[266] Graham Kalton and Howard Schuman.
The effect of the question on survey responses: A review.
Journal of the Royal Statistical Society. Series A (General), pages 42–73, 1982.

[267] Eser Kandogan.
Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions.
In *Proceedings of the IEEE Information Visualization Symposium*, volume 650, page 22. Citeseer, 2000.

[268] Oral Kaplan, Goshiro Yamamoto, Yasuhide Yoshitake, Takafumi Taketomi, Christian Sandor, and Hirokazu Kato.
In-situ visualization of pedaling forces on cycling training videos.
In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 000994–000999. IEEE, 2016.

[269] Niklas Karlsson, George Loewenstein, and Duane Seppi.
The ostrich effect: Selective attention to information.
Journal of Risk and uncertainty, 38(2):95–115, 2009.

[270] Uday S Karmarkar.
Subjectively weighted utility: A descriptive extension of the expected utility model.
Organizational behavior and human performance, 21(1):61–72, 1978.

[271] Matthew Kay and Jeffrey Heer.
Beyond weber's law: A second look at ranking visualizations of correlation.
IEEE transactions on visualization and computer graphics, 22(1):469–478, 2016.

[272] Daniel A Keim.
Pixel-oriented visualization techniques for exploring very large data bases.
Journal of Computational and Graphical Statistics, 5(1):58–77, 1996.

[273] Daniel A Keim and H-P Kriegel.
Visualization techniques for mining large databases: A comparison.
IEEE Transactions on knowledge and data engineering, 8(6):923–938, 1996.

[274] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler.
Challenges in visual data analysis.
In *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pages 9–16. IEEE, 2006.

[275] Boaz Keysar and Anne S Henly.
Speakers' overestimation of their effectiveness.
Psychological Science, 13(3):207–212, 2002.

[276] Azam Khan, Simon Breslav, Michael Glueck, and Kasper Hornbæk.
Benefits of visualization in the mammography problem.
International Journal of Human-Computer Studies, 83:94–113, 2015.

[277] Kris N Kirby and Daniel Gerlanc.
Bootes: An r package for bootstrap confidence intervals on effect sizes.
Behavior research methods, 45(4):905–927, 2013.

[278] Aniket Kittur, Ed H Chi, and Bongwon Suh.
Crowdsourcing user studies with mechanical turk.
In *Proc. of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[279] Joshua Klayman and Kaye Brown.
Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment.
Cognition, 49(1):97–122, 1993.

[280] Joshua Klayman, Jack B Soll, Claudia González-Vallejo, and Sema Barlas.
Overconfidence: It depends on how, what, and whom you ask.
Organizational behavior and human decision processes, 79(3):216–247, 1999.

[281] Alexander Klippel, Frank Hardisty, and Chris Weaver.
Star plots: How shape characteristics influence classification tasks.
Cartography and Geographic Information Science, 36(2):149–163, 2009.

[282] Jack L Knetsch.
The endowment effect and evidence of nonreversible indifference curves.
The american Economic review, 79(5):1277–1284, 1989.

[283] Ron Kohavi, Randal M Henne, and Dan Sommerfield.
Practical guide to controlled experiments on the web: listen to your customers not to the hippo.
In *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967. ACM, 2007.

[284] Artem Konev, Jürgen Waser, Bernhard Sadransky, Daniel Cornel, Rui AP Perdigao, Zsolt Horváth, and M Eduard Gröller.
Run watchers: Automatic simulation-based decision support in flood management.
IEEE transactions on visualization and computer graphics, 20(12):1873–1882, 2014.

[285] Asher Koriat, Morris Goldsmith, and Ainat Pansky.
Toward a psychology of memory accuracy.
Annual review of psychology, 51(1):481–537, 2000.

[286] Robert Kosara, Gerald N Sahling, and Helwig Hauser.
Linking scientific and information visualization with interactive 3d scatterplots.
2004.

[287] Justin Kruger.
Lake wobegon be gone! the" below-average effect" and the egocentric nature of comparative ability judgments.
Journal of personality and social psychology, 77(2):221, 1999.

[288] Justin Kruger and David Dunning.
Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.
Journal of personality and social psychology, 77(6):1121, 1999.

[289] Justin Kruger and Thomas Gilovich.
" naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias.
Journal of Personality and Social Psychology, 76(5):743, 1999.

[290] Martin Krzywinski and Naomi Altman.
Points of significance: error bars.
Nature methods, 10(10):921–922, 2013.

[291] X Kuang, H Zhang, S Zhao, and Michael J McGuffin.
Tracing tuples across dimensions: A comparison of scatterplots and parallel coordinate plots.
In *Computer Graphics Forum*, volume 31, pages 1365–1374. Wiley Online Library, 2012.

[292] Erica D Kuligowski and Steve MV Gwynne.
The need for behavioral theory in evacuation modeling.
Pedestrian and evacuation dynamics 2008, pages 721–732, 2010.

[293] Richard P Larrick, James N Morgan, and Richard E Nisbett.
Teaching the use of cost-benefit reasoning in everyday life.
Psychological Science, 1(6):362–370, 1990.

[294] Richard P Larrick, Richard E Nisbett, and James N Morgan.
Who uses the cost-benefit rules of choice? implications for the normative status of microeconomic theory.
Organizational Behavior and Human Decision Processes, 56(3):331–347, 1993.

[295] Tanya Latty and Madeleine Beekman.
Irrational decision-making in an amoeboid organism: transitivity and context-dependent preferences.
Proceedings of the Royal Society B: Biological Sciences, 278(1703):307, 2011.

[296] Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith.
Sketchstory: Telling more engaging stories with data through freeform sketching.
TVCG, 19(12):2416–2425, 2013.

[297] Michael D Lee, Rachel E Reilly, and Marcus E Butavicius.
An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data.
In *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24*, pages 1–10. Australian Computer Society, Inc., 2003.

[298] Darrin R Lehman and Richard E Nisbett.
A longitudinal study of the effects of undergraduate training on reasoning.
Developmental Psychology, 26(6):952, 1990.

[299] Dirk J Lehmann, Georgia Albuquerque, Martin Eisemann, Marcus Magnor, and Holger Theisel.
Selecting coherent and relevant plots in large scatterplot matrices.
In *Computer Graphics Forum*, volume 31, pages 1895–1908. Wiley Online Library, 2012.

[300] Stelios Lelis and Andrew Howes.
Informing decisions: how people use online rating information to make choices.
In *Proceedings of CHI 2011*, pages 2285–2294. ACM, 2011.

[301] Sylvain Lespinats and Michaël Aupetit.
Classimap: a supervised mapping technique for decision support.
In *Proceedings of the EuroVis Workshop on Visual Analytics using Multidimensional Projections*, pages 17–20, 2013.

[302] Alexander Lex, Marc Streit, Ernst Kruijff, and Dieter Schmalstieg.
Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context.
In *Visualization Symposium (PacificVis), 2010 IEEE Pacific*, pages 57–64. IEEE, 2010.

[303] Jing Li, Jean-Bernard Martens, and Jarke J Van Wijk.
Judging correlation from scatterplots and parallel coordinate plots.
Information Visualization, 9(1):13–30, 2010.

[304] Sarah Lichtenstein and Baruch Fischhoff.
Do those who know more also know more about how much they know?
Organizational behavior and human performance, 20(2):159–183, 1977.

[305] Sarah Lichtenstein and Paul Slovic.
Reversals of preference between bids and choices in gambling decisions.
Journal of experimental psychology, 89(1):46, 1971.

[306] D Stephen Lindsay and Marcia K Johnson.
The eyewitness suggestibility effect and memory for source.
Memory & Cognition, 17(3):349–358, 1989.

[307] Ran Liu, Tiffany Chao, Catherine Plaisant, and Ben Shneiderman.
Manylists: product comparison tool using spatial layouts with animated transitions.
University of Maryland Technical Report, 2012.

[308] George Loewenstein.
Hot-cold empathy gaps and medical decision making.
Health Psychology, 24(4S):S49, 2005.

[309] Elizabeth F Loftus and John C Palmer.
Reconstruction of automobile destruction: An example of the interaction between language and memory.
Journal of verbal learning and verbal behavior, 13(5):585–589, 1974.

[310] Graham Loomes and Robert Sugden.
Some implications of a more general form of regret theory.
Journal of Economic Theory, 41(2):270–287, 1987.

[311] Hela Ltifi, Emna Ben Mohamed, and Mounir ben Ayed.
Interactive visual knowledge discovery from data-based temporal decision support system.
Information Visualization, 15(1):31–50, 2016.

[312] Martin Luboschik, Martin Röhlig, Günther Kundt, Oliver Stachs, Sabine Peschel, Andrey Zhivov, Rudolf F Guthoff, Karsten Winter, and Heidrun Schumann.
Supporting an early detection of diabetic neuropathy by visual analytics.
In *Proceedings of the EuroVis Workshop on Visual Analytics (EuroVA)*, 2014.

[313] Kwan-Liu Ma, Isaac Liao, Jennifer Frazier, Helwig Hauser, and Helen-Nicole Kostis.
Scientific storytelling using visualization.
IEEE Computer Graphics and Applications, 32(1):12–19, 2012.

[314] C Neil Macrae and Galen V Bodenhausen.
Social cognition: Thinking categorically about others.
Annual review of psychology, 51(1):93–120, 2000.

[315] Michael J Mahoney.
Publication prejudices: An experimental study of confirmatory bias in the peer review system.
Cognitive therapy and research, 1(2):161–175, 1977.

[316] Selin A. Malkoc, William Hedgcock, and Steve Hoeffler.
Between a rock and a hard place: The failure of the attraction effect among unattractive alternatives.
Journal of Consumer Psychology, 23(3):317 – 329, 2013.

[317] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch.
Human performance consequences of automated decision aids: The impact of degree of automation and system experience.
Journal of Cognitive Engineering and Decision Making, 6(1):57–87, 2012.

[318] Rosa Mikeal Martey, Adrienne Shaw, Jennifer Stromer-Galley, Kate Kenski, Benjamin A Clegg, James E Folkestad, Emilie T Saulnier, and Tomek Strzalkowski.
Testing the power of game lessons: The effects of art and narrative on reducing cognitive biases.
In *DiGRA*, 2014.

[319] Winter Mason and Duncan J Watts.
Financial incentives and the performance of crowds.
ACM SigKDD Explorations Newsletter, 11(2):100–108, 2010.

[320] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice.
The effect of visual appearance on the performance of continuous sliders and visual analogue scales.
In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5421–5432. ACM, 2016.

[321] Mara Mather and Laura L Carstensen.
Aging and motivated cognition: The positivity effect in attention and memory.
Trends in cognitive sciences, 9(10):496–502, 2005.

[322] Mara Mather and Marcia K Johnson.
Choice-supportive source monitoring: Do our decisions seem better to us as we age?
Psychology and aging, 15(4):596, 2000.

[323] Manos Matsaganis and Chrysa Leventi.
Poverty and inequality during the great recession in greece.
Political Studies Review, 12(2):209–223, 2014.

[324] John D Mayer, Laura J McCormick, and Sara E Strong.
Mood-congruent memory and natural mood: New evidence.
Personality and Social Psychology Bulletin, 21(7):736–746, 1995.

[325] Dawn M McBride and Barbara Anne Dosher.
A comparison of conscious and automatic memory processes for picture and word stimuli: A process dissociation analysis.
Consciousness and cognition, 11(3):423–460, 2002.

[326] Gilles O. McDaniel, Mark A.; Einstein.
Bizarre imagery as an effective memory aid: The importance of distinctiveness.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 1986.
doi: 10.1037/0278-7393.12.1.54.
URL `http://gen.lib.rus.ec/scimag/index.php?s=10.1037/0278-7393.12.1.54`.

[327] Mark A McDaniel, Gilles O Einstein, Edward L DeLosh, Cindi P May, and Paul Brady.
The bizarreness effect: it's not surprising, it's complex.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(2):422, 1995.

[328] Mark A McDaniel, Janis L Anderson, Mary H Derbish, and Nova Morrisette.
Testing the testing effect in the classroom.
European Journal of Cognitive Psychology, 19(4-5):494–513, 2007.

[329] Kevin T McDonnell and Klaus Mueller.
Illustrative parallel coordinates.
In *Computer Graphics Forum*, volume 27, pages 1031–1038. Wiley Online Library, 2008.

[330] Matthew S. McGlone and Jessica Tofighbakhsh.
Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms.
Psychological Science, 11(5):424–428, 2000.
doi: 10.1111/1467-9280.00282.

[331] Paul E Meehl.
Why i do not attend case conferences.
Psychodiagnosis: selected papers, pages 225–302, 1973.

[332] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete.
Assessing the effect of visualizations on bayesian reasoning through crowdsourcing.
IEEE Transactions on Visualization and Computer Graphics, 18(12):2536–2545, 2012.

[333] MA Migut, Jan C van Gemert, and Marcel Worring.
Interactive decision making using dissimilarity to visually represented prototypes.
In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 141–149. IEEE, 2011.

[334] Stanley Milgram.
Behavioral study of obedience.
The Journal of abnormal and social psychology, 67(4):371, 1963.

[335] Dale T Miller and Stephen A Norman.
Actor-observer differences in perceptions of effective control.
Journal of Personality and Social Psychology, 31(3):503, 1975.

[336] Sanjay Mishra, UN Umesh, and Donald E Stem Jr.
Antecedents of the attraction effect: An information-processing approach.
Journal of Marketing Research, pages 331–349, 1993.

[337] Gabriel Mistelbauer, Arnold Köchl, Rudiger Schernthaner, Ivan Baclija, Rüdiger Schernthaner, Stefan Bruckner, Milos Sramek, and Meister Eduard Gröller.
Smart super views—a knowledge-assisted interface for medical visualization.
In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 163–172. IEEE, 2012.

[338] Terence R Mitchell, Leigh Thompson, Erika Peterson, and Randy Cronk.
Temporal adjustments in the evaluation of events: The "rosy view".
Journal of Experimental Social Psychology, 33(4):421–448, 1997.

[339] Benoit Monin and Dale T Miller.
Moral credentials and the expression of prejudice.
Journal of personality and social psychology, 81(1):33, 2001.

[340] Carey K Morewedge and Colleen E Giblin.
Explanations of the endowment effect: an integrative review.
Trends in cognitive sciences, 19(6):339–348, 2015.

[341] Carey K Morewedge, Haewon Yoon, Irene Scopelliti, Carl W Symborski, James H Korris, and Karim S Kassam.
Debiasing decisions: Improved decision making with a single training intervention.
Policy Insights from the Behavioral and Brain Sciences, 2(1):129–140, 2015.

[342] Manfred Mortell, Hanan H Balkhy, Elias B Tannous, and Mei Thiee Jong.
Physician 'defiance'towards hand hygiene compliance: Is there a theory–practice–ethics gap?
Journal of the Saudi Heart Association, 25(3):203–208, 2013.

[343] John Morton, Robert G Crowder, and Harvey A Prussin.
Experiments with the stimulus suffix effect.
1971.

[344] Serge Moscovici and Marisa Zavalloni.
The group as a polarizer of attitudes.
Journal of personality and social psychology, 12(2):125, 1969.

[345] Sumitava Mukherjee and Narayanan Srinivasan.
Attention in preferential choice.
Prog. Brain Res, 202:117–134, 2013.

[346] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge.
Survey research in hci.
In *Ways of Knowing in HCI*, pages 229–266. Springer, 2014.

[347] Tamara Munzner.
Visualization Analysis and Design.
CRC Press, 2014.

[348] Bennet B Murdock Jr.
The serial position effect of free recall.
Journal of experimental psychology, 64(5):482, 1962.

[349] Thomas Mussweiler, Fritz Strack, and Tim Pfeiffer.
Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility.
Personality and Social Psychology Bulletin, 26(9):1142–1150, 2000.

[350] Richard Nadeau, Edouard Cloutier, and J-H Guay.
New evidence about the existence of a bandwagon effect in the opinion formation process.
International Political Science Review, 14(2):203–213, 1993.

[351] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman.
Deliberation increases the wisdom of crowds.
arXiv preprint arXiv:1703.00045, 2017.

[352] Raymond S Nickerson.
Confirmation bias: A ubiquitous phenomenon in many guises.
Review of general psychology, 2(2):175, 1998.

[353] Richard E Nisbett and Timothy D Wilson.
The halo effect: Evidence for unconscious alteration of judgments.
Journal of personality and social psychology, 35(4):250, 1977.

[354] Loran F Nordgren, Frenk van Harreveld, and Joop van der Pligt.
The restraint bias: How the illusion of self-restraint promotes impulsive behavior.
Psychological Science, 20(12):1523–1528, 2009.

[355] Michael I Norton, Daniel Mochon, and Dan Ariely.
The ikea effect: When labor leads to love.
Journal of Consumer Psychology, 22(3):453–460, 2012.

[356] Brendan Nyhan and Jason Reifler.
When corrections fail: The persistence of political misperceptions.
Political Behavior, 32(2):303–330, 2010.

[357] Edward J O'brien and Jerome L Myers.
When comprehension difficulty improves memory for text.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 11(1):12, 1985.

[358] Yigang Pan Sue O'Curry and Robert Pitts.
The attraction effect and political choice in two elections.
Journal of Consumer Psychology, 4(1):85–101, 1995.

[359] Steffen Oeltze, Helmut Doleisch, Helwig Hauser, Philipp Muigg, and Bernhard Preim.
Interactive visual analysis of perfusion data.
IEEE transactions on visualization and computer graphics, 13(6):1392–1399, 2007.

[360] Mehrnoosh Oghbaie, Michael J Pennock, and William B Rouse.
Understanding the efficacy of interactive visualization for decision making for complex systems.
In *Systems Conference (SysCon), 2016 Annual IEEE*, pages 1–6. IEEE, 2016.

[361] Maria Cristina Ferreira de Oliveira and Haim Levkowitz.
From visual data exploration to visual data mining: a survey.
IEEE Transactions on Visualization and Computer Graphics, 9(3):378–394, 2003.

[362] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko.
Instructional manipulation checks: Detecting satisficing to increase statistical power.
JESP, 45(4):867–872, 2009.

[363] Rickard O Ottosson, Per E Engström, David Sjöström, Claus F Behrens, Anna Karlsson, Tommy Knöös, and Crister Ceberg.
The feasibility of using pareto fronts for comparison of treatment planning systems and delivery techniques.
Acta Oncologica, 48(2):233–237, 2009.

[364] Stephan Pajer, Marc Streit, Thomas Torsney-Weir, Florian Spechtenhauser, Torsten Muller, and Harald Piringer.
Weightlifter: Visual weight space exploration for multi-criteria decision making.
IEEE transactions on visualization and computer graphics, 23(1):611, 2017.

[365] Pietro Panzarasa, Nicholas R Jennings, and Timothy J Norman.
Formalizing collaborative decision-making and practical reasoning in multi-agent systems.
Journal of logic and computation, 12(1):55–117, 2002.

[366] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis.
Running experiments on amazon mechanical turk.
Judgment and Decision making, 5(5):411–419, 2010.

[367] Bernadette Park and Myron Rothbart.
Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members.
Journal of Personality and Social Psychology, 42(6):1051, 1982.

[368] Amanda Parker, Edward Wilding, and Colin Akerman.
The von restorff effect in visual object recognition memory in humans and monkeys: The role of frontal/perirhinal interaction.
Journal of Cognitive Neuroscience, 10(6):691–703, 1998.

[369] Giulia Pastorella.
Technocratic governments in europe: getting the critique right.
Political Studies, 64(4):948–965, 2016.

[370] Helen M Paterson, Richard I Kemp, and Joseph P Forgas.
Co-witnesses, confederates, and conformity: Effects of discussion and delay on eyewitness memory.
Psychiatry, Psychology and Law, 16(sup1):S112–S124, 2009.

[371] John W Payne.
Alternative approaches to decision making under risk.
Psychological Bulletin, 80(6):439–453, 1973.

[372] John W Payne, James R Bettman, and Eric J Johnson.
The adaptive decision maker.
Cambridge University Press, 1993.

[373] K Peason.
On lines and planes of closest fit to systems of point in space.
Philosophical Magazine, 2(11):559–572, 1901.

[374] Charles Perin, Pierre Dragicevic, and Jean-Daniel Fekete.
Revisiting bertin matrices: New interactions for crafting tabular visualizations.
IEEE transactions on visualization and computer graphics, 20(12):2082–2091, 2014.

[375] Jonathan C. Pettibone.
Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice.
Judgment and Decision Making, 7(4):513–523, 2012.

[376] Jonathan C Pettibone and Douglas H Wedell.
Examining models of nondominated decoy effects across judgment and choice.
Organizational Behavior and Human Decision Processes, 81(2):300 – 328, 2000.

[377] Thomas F Pettigrew.
The ultimate attribution error: Extending allport's cognitive analysis of prejudice.
Personality and social psychology bulletin, 5(4):461–476, 1979.

[378] Lawrence D Phillips and Ward Edwards.
Conservatism in a simple probability inference task.
Journal of experimental psychology, 72(3):346, 1966.

[379] Catherine Plaisant.
The challenge of information visualization evaluation.
In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

[380] Scott Plous.
The psychology of judgment and decision making.
Mcgraw-Hill Book Company, 1993.

[381] Margit Pohl.
Cognitive biases in visual analytics – a critical reflection. in press.
In *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations. IEEE VIS.*, 2017.

[382] Rüdiger Pohl.
Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory.
Psychology Press, 2004.

[383] J Poppenk, G Walia, AR McIntosh, MF Joanisse, D Klein, and S Köhler.
Why is the meaning of a sentence better remembered than its form? an fmri study on the role of novelty-encoding processes.
Hippocampus, 18(9):909–918, 2008.

[384] Zachary Pousman, John Stasko, and Michael Mateas.
Casual information visualization: Depictions of data in everyday life.
IEEE transactions on visualization and computer graphics, 13(6), 2007.

[385] Dražen Prelec, H Sebastian Seung, and John McCoy.
A solution to the single-question crowd wisdom problem.
Nature, 541(7638):532–535, 2017.

[386] Emily Pronin and Matthew B Kugler.
Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot.
Journal of Experimental Social Psychology, 43(4):565–578, 2007.

[387] Emily Pronin and Lee Ross.
Temporal differences in trait self-ascription: when the self is seen as an other.
Journal of personality and social psychology, 90(2):197, 2006.

[388] Emily Pronin, Justin Kruger, Kenneth Savtisky, and Lee Ross.
You don't know me, but i know you: The illusion of asymmetric insight.
Journal of Personality and Social Psychology, 81(4):639, 2001.

[389] Pearl Pu and Boi Faltings.
Enriching buyers' experiences: the smartclient approach.
In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 289–296. ACM, 2000.

[390] Priya Raghubir and Joydeep Srivastava.
The denomination effect.
Journal of Consumer Research, 36(4):701–713, 2009.

[391] Ramana Rao and Stuart K Card.
The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322. ACM, 1994.

[392] Srinivasan Ratneshwar, Allan D Shocker, and David W Stewart.
Toward understanding the attraction effect: The implications of product stimulus meaningfulness and familiarity.
J. of Consumer Research, pages 520–533, 1987.

[393] William E Remus and Jeffrey E Kottemann.
Toward intelligent decision support systems: An artificially intelligent statistician.
MIS Quarterly, pages 403–418, 1986.

[394] Ronald A Rensink and Gideon Baldridge.
The perception of correlation in scatterplots.
In *Computer Graphics Forum*, volume 29, pages 1203–1210. Wiley Online Library, 2010.

[395] Valerie F Reyna, Christina F Chick, Jonathan C Corbin, and Andrew N Hsia.
Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students.
Psychological science, 25(1):76–84, 2014.

[396] Patrick Riehmann, Jens Opolka, and Bernd Froehlich.
The product explorer: decision making with ease.
In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 423–432. ACM, 2012.

[397] Ilana Ritov and Jonathan Baron.
Reluctance to vaccinate: Omission bias and ambiguity.
Journal of Behavioral Decision Making, 3(4):263–277, 1990.

[398] Ralph Rosnow and Robert Rosenthal.
People studying people: Artifacts and ethics in behavioral research.
WH Freeman, 1997.

[399] Lee Ross and Constance Stillinger.
Barriers to conflict resolution.
Negotiation Journal, 7(4):389–404, 1991.

[400] Lee Ross, David Greene, and Pamela House.
The "false consensus effect": An egocentric bias in social perception and attribution processes.
Journal of experimental social psychology, 13(3):279–301, 1977.

[401] Michael Ross, Fiore Sicoly, et al.
Egocentric biases in availability and attribution.
Journal of personality and social psychology, 37(3):322–336, 1979.

[402] Steven F Roth and Joe Mattis.
Data characterization for intelligent graphics presentation.
In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–200. ACM, 1990.

[403] Joanna Różycka-Tran, Paweł Boski, and Bogdan Wojciszke.
Belief in a zero-sum game as a social axiom: A 37-nation study.
Journal of Cross-Cultural Psychology, 46(4):525–548, 2015.

[404] Laurie A Rudman and Stephanie A Goodwin.
Gender differences in automatic in-group bias: Why do women like women more than men like men?
Journal of personality and social psychology, 87(4):494, 2004.

[405] Stephen Rudolph, Anya Savikhin, and David S Ebert.
Finvis: Applied visual analytics for personal financial planning.
In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 195–202. IEEE, 2009.

[406] J Edward Russo and Barbara A Dosher.
Strategies for multiattribute binary choice.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(4):676, 1983.

[407] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim.
The role of uncertainty, awareness, and trust in visual analytics.
IEEE transactions on visualization and computer graphics, 22(1):240–249, 2016.

[408] William Samuelson and Richard Zeckhauser.
Status quo bias in decision making.
Journal of risk and uncertainty, 1(1):7–59, 1988.

[409] Ignacio Sánchez-Cuenca.
From a deficit of democracy to a technocratic order: The postcrisis debate on europe.
Annual Review of Political Science, (0), 2017.

[410] Lawrence J Sanna and Norbert Schwarz.
Integrating temporal biases: The interplay of focal thoughts and accessibility experiences.
Psychological science, 15(7):474–481, 2004.

[411] Lawrence J Sanna, Norbert Schwarz, and Shevaun L Stocker.
When debiasing backfires: accessible content and accessibility experiences in debiasing hindsight.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3):497, 2002.

[412] Purvi Saraiya, Chris North, and Karen Duca.
An insight-based methodology for evaluating bioinformatics visualizations.
TVCG, 11(4):443–456, July 2005.
ISSN 1077-2626.

[413] Arvind Satyanarayan and Jeffrey Heer.
Authoring narrative visualizations with ellipsis.
In *Proc. of the 16th Eurographics Conference on Visualization*, EuroVis '14, pages 361–370, Aire-la-Ville, Switzerland, Switzerland, 2014. Eurographics Association.

[414] J Savage Leonard.
The foundations of statistics.
NY, John Wiley, pages 188–190, 1954.

[415] Anya Savikhin, Hon Cheong Lam, Brian Fisher, and David S Ebert.
An experimental study of financial portfolio selection with visual analytics for decision support.
In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[416] Kenneth Savitsky and Thomas Gilovich.
The illusion of transparency and the alleviation of speech anxiety.
Journal of experimental social psychology, 39(6):618–625, 2003.

[417] David A Schkade and Daniel Kahneman.
Does living in california make people happy? a focusing illusion in judgments of life satisfaction.
Psychological Science, 9(5):340–346, 1998.

[418] Ulrich Schmidt and Stefan Traub.
An experimental test of loss aversion.
Journal of Risk and Uncertainty, 25(3):233–249, 2002.

[419] Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann.
A design space of visualization tasks.
IEEE Transactions on Visualization and Computer Graphics, 19(12):2366–2375, 2013.

[420] M Sedlmair, M Brehmer, S Ingram, and T Munzner.
Dimensionality reduction in the wild: Gaps and guidance.
Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, 2012.

[421] Edward Segel and Jeffrey Heer.
Narrative visualization: Telling stories with data.
Visualization and Computer Graphics, IEEE Transactions on, 16(6):1139–1148, 2010.

[422] Sankar Sen.
Knowledge, information mode, and the attraction effect.
Journal of Consumer Research, 25(1):64–77, 1998.

[423] Eldar Shafir, Peter Diamond, and Amos Tversky.
Money illusion.
The Quarterly Journal of Economics, 112(2):341–374, 1997.

[424] Sharoni Shafir, Tom A. Waite, and Brian H. Smith.
Context-dependent violations of rational choice in honeybees (apis mellifera) and gray jays (perisoreus canadensis).
Behavioral Ecology and Sociobiology, 51(2):180–187, 2002.

[425] Tali Sharot, Alison M Riccardi, Candace M Raio, and Elizabeth A Phelps.
Neural mechanisms mediating optimism bias.
Nature, 450(7166):102–105, 2007.

[426] Kelly G Shaver.
Defensive attribution: Effects of severity and relevance on the responsibility assigned for an accident.
Journal of Personality and Social Psychology, 14(2):101, 1970.

[427] Angela Shen-Hsieh and Mark Schindl.
Data visualization for strategic decision making.
In *Case Studies of the CHI2002*, pages 1–17. ACM, 2002.

[428] Ben Shneiderman and Catherine Plaisant.
Treemaps for space-constrained visualization of hierarchies.
1998.

[429] Harri Siirtola.
Interaction with the reorderable matrix.
In *Information Visualization, 1999. Proceedings. 1999 IEEE International Conference on*, pages 272–277. IEEE, 1999.

[430] Harri Siirtola and Kari-Jouko Räihä.
Interacting with parallel coordinates.
Interacting with Computers, 18(6):1278–1309, 2006.

[431] Herbert A Simon.
Rational choice and the structure of the environment.
Psychological review, 63(2):129, 1956.

[432] Herbert A Simon.
The new science of management decision.
1960.

[433] Itamar Simonson.
Choice based on reasons: The case of attraction and compromise effects.
Journal of Consumer Research, 16(2):158–174, 1989.

[434] Itamar Simonson.
Vices and virtues of misguided replications: The case of asymmetric dominance.
J. of Marketing Research, 51(4):514–519, 2014.
doi: 10.1509/jmr.14.0093.

[435] Itamar Simonson and Amos Tversky.
Choice in context: Tradeoff contrast and extremeness aversion.
Journal of Marketing Research, 29:281–295, 1992.

[436] Norman J Slamecka.
An examination of trace storage in free recall.
Journal of experimental psychology, 76(4p1):504, 1968.

[437] Norman J Slamecka and Peter Graf.
The generation effect: Delineation of a phenomenon.
Journal of experimental Psychology: Human learning and Memory, 4(6):592, 1978.

[438] Jack B Soll and Richard P Larrick.
Strategies for revising judgment: How (and how well) people use others' opinions.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(3):780, 2009.

[439] Johannes Sorger, Thomas Ortner, Christian Luksch, Michael Schwarzler, Eduard Groller, and Harald Piringer.
Litevis: Integrated visualization for simulation-based decision support in lighting design.
Visualization and Computer Graphics, IEEE Transactions on, 22(1):290–299, 2016.

[440] Betsy Sparrow, Jenny Liu, and Daniel M Wegner.
Google effects on memory: Cognitive consequences of having information at our fingertips.
science, 333(6043):776–778, 2011.

[441] Ian Spence.
No humble pie: The origins and usage of a statistical chart.
Journal of Educational and Behavioral Statistics, 30(4):353–368, 2005.

[442] Robert Spence and Lisa Tweedie.
The attribute explorer: information synthesis via exploration.
Interacting with Computers, 11(2):137–146, 1998.

[443] Michael Spenke, Christian Beilken, and Thomas Berlage.
Focus: the interactive table for product comparison and selection.
In *Proceedings of the 9th annual ACM symposium on User interface software and technology*, pages 41–50. ACM, 1996.

[444] Mark Spranca, Elisa Minsk, and Jonathan Baron.
Omission and commission in judgment and choice.
Journal of experimental social psychology, 27(1):76–105, 1991.

[445] Rebecca R Springmeyer, Meera M Blattner, and Nelson L Max.
A characterization of the scientific data analysis process.
In *Proceedings of the 3rd conference on Visualization'92*, pages 235–242. IEEE Computer Society Press, 1992.

[446] Diederik A Stapel, Leonard L Martin, and Norbert Schwarz.
The smell of bias: What instigates correction processes in social judgments?
Personality and Social Psychology Bulletin, 24(8):797–806, 1998.

[447] Barry M Staw.
Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action.
Organizational behavior and human performance, 16(1):27–44, 1976.

[448] Charles D Stolper, Adam Perer, and David Gotz.
Progressive visual analytics: User-driven visual exploration of in-progress analytics.
IEEE Transactions on Visualization and Computer Graphics, 20(12):1653–1662, 2014.

[449] Sergej Stoppel and Stefan Bruckner.
Vol 2 velle: Printable interactive volume visualization.
IEEE transactions on visualization and computer graphics, 23(1):861–870, 2017.

[450] Alexander Streit, Binh Pham, and Ross Brown.
A spreadsheet approach to facilitate visualization of uncertainty in information.
IEEE Transactions on Visualization and Computer Graphics, 14(1):61–72, 2008.

[451] Hannah Summerfelt, Louis Lippman, and Ira E Hyman Jr.
The effect of humor on memory: Constrained by the pun.
The Journal of General Psychology, 137(4):376–394, 2010.

[452] Cass R Sunstein.
Probability neglect: Emotions, worst cases, and law.
The Yale Law Journal, 112(1):61–107, 2002.

[453] James Surowiecki.
The wisdom of crowds.
Anchor, 2005.

[454] Alistair Sutcliffe, Oscar de Bruijn, Sarah Thew, Iain Buchan, Paul Jarvis, John McNaught, and Rob Procter.
Developing visualization-based decision support tools for epidemiology.
Information Visualization, 13(1):3–17, 2014.

[455] Ola Svenson.
Decisions among time saving options: When intuition is strong and wrong.
Acta Psychologica, 127(2):501–509, 2008.

[456] Justin Talbot, Vidya Setlur, and Anushka Anand.
Four experiments on the perception of bar charts.
IEEE transactions on visualization and computer graphics, 20(12):2152–2160, 2014.

[457] James W Tanaka, Markus Kiefer, and Cindy M Bukach.
A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study.
Cognition, 93(1):B1–B9, 2004.

[458] Donald M Taylor and Janet R Doria.
Self-serving and group-serving bias in attribution.
The Journal of Social Psychology, 113(2):201–211, 1981.

[459] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger.
The perfect search engine is not enough: a study of orienteering behavior in directed search.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422. ACM, 2004.

[460] Richard Thaler.
Some empirical evidence on dynamic inconsistency.
Economics letters, 8(3):201–207, 1981.

[461] Richard H Thaler, Amos Tversky, Daniel Kahneman, and Alan Schwartz.
The effect of myopia and loss aversion on risk taking: An experimental test.
The Quarterly Journal of Economics, 112(2):647–661, 1997.

[462] Suzanne C Thompson.
Illusions of control: How we overestimate our personal influence.
Current Directions in Psychological Science, 8(6):187–190, 1999.

[463] Jerome Tobacyk, Gary Milford, Thomas Springer, and Zofia Tobacyk.
Paranormal beliefs and the barnum effect.
Journal of Personality Assessment, 52(4):737–739, 1988.

[464] Melanie Tory and Torsten Moller.
Human factors in visualization research.
IEEE transactions on visualization and computer graphics, 10(1):72–84, 2004.

[465] J Gregory Trafton, Susan S Kirschenbaum, Ted L Tsui, Robert T Miyamoto, James A Ballas, and Paula D Raymond.
Turning pictures into numbers: extracting and generating information from complex visualizations.
International Journal of Human-Computer Studies, 53(5):827–850, 2000.

[466] Evangelos Triantaphyllou.
Multi-criteria decision making methods: a comparative study, volume 44.
Springer Science & Business Media, 2013.

[467] Jennifer S Trueblood, Scott D Brown, Andrew Heathcote, and Jerome R Busemeyer.
Not just for consumers context effects are fundamental to decision making.
Psychological science, 24(6):901–908, 2013.

[468] JenniferS. Trueblood.
Multialternative context effects obtained using an inference task.
Psychonomic Bulletin and Review, 19(5):962–968, 2012.

[469] John W Tukey.
Exploratory data analysis.
1977.

[470] Endel Tulving.
Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible.
American Scientist, 62(1):74–82, 1974.

[471] A. Tversky.
Elimination by aspects: A theory of choice.
Psychological Review, 79:281–299, 1972.

[472] Amos Tversky.
Intransitivity of preferences.
Psychological Review, pages 31–248, 1969.

[473] Amos Tversky and Daniel Kahneman.
Availability: A heuristic for judging frequency and probability.
Cognitive psychology, 5(2):207–232, 1973.

[474] Amos Tversky and Daniel Kahneman.
Judgment under uncertainty: Heuristics and biases.
In *Utility, probability, and human decision making*, pages 141–162. Springer, 1975.

[475] Amos Tversky and Daniel Kahneman.
The framing of decisions and the psychology of choice.
Science, 1981.

[476] Amos Tversky and Daniel Kahneman.
Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.
Psychological review, 90(4):293, 1983.

[477] Amos Tversky and Daniel Kahneman.
Rational choice and the framing of decisions.
Journal of business, pages S251–S278, 1986.

[478] Amos Tversky and Derek J Koehler.
Support theory: A nonextensional representation of subjective probability.
Psychological review, 101(4):547, 1994.

[479] Lisa Tweedie and Robert Spence.
The prosection matrix: a tool to support the interactive exploration of statistical models and data.
Computational Statistics, 13(1):65–76, 1998.

[480] JoNell A Usher and Ulric Neisser.
Childhood amnesia and the beginnings of memory for four early life events.
Journal of Experimental Psychology: General, 122(2):155, 1993.

[481] Andre Calero Valdez, Martina Ziefle, and Michael Sedlmair.
Priming and anchoring effects in visualizations.
IEEE Transactions on Visualization and Computer Graphics, 2017.

[482] Venu Vasudevan.
Supporting high-bandwidth navigation in object-bases.
In *Data Engineering, 1994. Proceedings. 10th International Conference*, pages 294–301. IEEE, 1994.

[483] Mark Velasquez and Patrick T Hester.
An analysis of multi-criteria decision making methods.
International Journal of Operations Research, 10(2):56–66, 2013.

[484] Susannah Verney and Anna Bosco.
Living parallel lives: Italy and greece in an age of austerity.
South European Society and Politics, 18(4):397–426, 2013.

[485] Christophe Viau, Michael J McGuffin, Yves Chiricota, and Igor Jurisica.
The flowvizmenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration.
IEEE Transactions on Visualization and Computer Graphics, 16(6):1100–1108.

[486] Fernanda B Viégas, Danah Boyd, David H Nguyen, Jeffrey Potter, and Judith Donath.
Digital artifacts for remembering and storytelling: Posthistory and social network fragments.
In *System Sciences, 2004. Proc. of the 37th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2004.

[487] Fernanda B Viégas, Scott Golder, and Judith Donath.
Visualizing email content: portraying relationships from conversational histories.
In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.

[488] John Von Neumann and Oskar Morgenstern.
Theory of games and economic behavior.
Princeton university press, 2007.

[489] Joel L Voss, Kara D Federmeier, and Ken A Paller.
The potato chip really does look like elvis! neural hallmarks of conceptual processing associated with finding novel shapes subjectively meaningful.
Cerebral Cortex, 22(10):2354–2364, 2011.

[490] Willem A Wagenaar and Gideon B Keren.
Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people.
Organizational Behavior and Human Decision Processes, 36(3):406–416, 1985.

[491] Drew Walker and Edward Vul.
Hierarchical encoding makes individuals in a group seem more attractive.
Psychological Science, 25(1):230–235, 2014.

[492] W Richard Walker and John J Skowronski.
The fading affect bias: But what the hell is it for?
Applied Cognitive Psychology, 23(8):1122–1136, 2009.

[493] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert.
Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics.
In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[494] Emily Wall, LM Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert.
Four perspectives on human bias in visual analytics. in press.
In *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations. IEEE VIS.*, 2017.

[495] Colin Ware.
Information visualization: perception for design.
Elsevier, 2012.

[496] Peter C Wason.
On the failure to eliminate hypotheses in a conceptual task.
Quarterly journal of experimental psychology, 12(3):129–140, 1960.

[497] Martin Wattenberg.
Baby names, visualization, and social data analysis.
In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 1–7. IEEE, 2005.

[498] Adam Waytz, John Cacioppo, and Nicholas Epley.
Who sees human? the stability and importance of individual differences in anthropomorphism.
Perspectives on Psychological Science, 5(3):219–232, 2010.

[499] Martin Weber and Colin F Camerer.
The disposition effect in securities trading: An experimental analysis.
Journal of Economic Behavior & Organization, 33(2):167–184, 1998.

[500] Douglas H Wedell.
Distinguishing among models of contextually induced preference reversals.
Journal of Experimental Psychology: Learning, Memory, and Cognition, 17(4):767, 1991.

[501] Douglas H. Wedell and Jonathan C. Pettibone.
Using judgments to understand decoy effects in choice.
Organizational Behavior and Human Decision Processes, 67(3):326 – 344, 1996.

[502] Rainer Wegenkittl, Helwig Loffelmann, and Eduard Groller.
Visualizing the behaviour of higher dimensional dynamical systems.
In *Visualization'97., Proceedings*, pages 119–125. IEEE, 1997.

[503] Peter Wegner.
Why interaction is more powerful than algorithms.
Communications of the ACM, 40(5):80–91, 1997.

[504] Stephen Wehrend and Clayton Lewis.
A problem-oriented classification of visualization techniques.
In *Proc. of Visualization'90*, pages 139–143. IEEE Computer Society Press, 1990.

[505] Neil D Weinstein.
Unrealistic optimism about future life events.
Journal of personality and social psychology, 39(5):806, 1980.

[506] Richard F West, Russell J Meserve, and Keith E Stanovich.
Cognitive sophistication does not attenuate the bias blind spot.
Journal of personality and social psychology, 103(3):506, 2012.

[507] Bernard E Whitley Jr and Irene Hanson Frieze.
Measuring causal attributions for success and failure: A meta-analysis of the effects of question-wording style.
Basic and Applied Social Psychology, 7(1):35–51, 1986.

[508] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja.
Graphical inference for infovis.
IEEE Transactions on Visualization and Computer Graphics, 16(6):973–979, 2010.

[509] Jarke J Wijk and Robert Liere.
Hyperslice visualization of scalar functions of many variables.
1994.

[510] Wikipedia.
List of cognitive biases — wikipedia, the free encyclopedia, 2017.
URL https://en.wikipedia.org/w/index.php?title=List_of_cognitive_biases&oldid=791032058.
[Online; accessed 23-July-2017].

[511] Wikipedia.
Open data — wikipedia, the free encyclopedia, 2017.
URL https://en.wikipedia.org/wiki/Open_data.
[Online; accessed 23-July-2017].

[512] Leland Wilkinson and Michael Friendly.
The history of the cluster heat map.
The American Statistician, 2012.

[513] Christopher Williamson and Ben Shneiderman.
The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system.
In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–346. ACM, 1992.

[514] Kent Wittenburg, Tom Lanning, Michael Heinrichs, and Michael Stanton.
Parallel bargrams for consumer-based information exploration and choice.
In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 51–60. ACM, 2001.

[515] Pak Chung Wong and R Daniel Bergeron.
30 years of multidimensional multivariate visualization.
In *Scientific Visualization*, pages 3–33, 1994.

[516] Jo Wood, Donia Badawood, Jason Dykes, and Aidan Slingsby.
Ballotmaps: Detecting name bias in alphabetically ordered ballot papers.
Visualization and Computer Graphics, IEEE Transactions on, 17(12):2384–2391, 2011.

[517] Peter Wright.
Consumer choice strategies: Simplifying vs. optimizing.
Journal of Marketing Research, pages 60–67, 1975.

[518] William Wright.
Information animation applications in the capital markets.
In *Information Visualization, 1995. Proceedings.*, pages 19–25. IEEE, 1995.

[519] Cindy Xiong, van Lisanne Weelden†, and Steven Franconeri.
The curse of knowledge in visual data communication. in press.
In *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations. IEEE VIS.*, 2017.

[520] Yangqing Xu and Steven L Franconeri.
The head of the table: Marking the "front" of an object is tightly linked with selection.
Journal of Neuroscience, 32(4):1408–1412, 2012.

[521] Mehmet Adil Yalçına, Niklas Elmqvistb, and Benjamin B Bedersona.
Evaluating multi-column bar charts and treemaps for dense visualization of sorted numeric data.
2015.

[522] Jing Yang, Wei Peng, Matthew O Ward, and Elke A Rundensteiner.
Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets.
In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 105–112. IEEE, 2003.

[523] Sybil Yang and Michael Lynn.
More evidence challenging the robustness and usefulness of the attraction effect.
Journal of Marketing Research, 51(4):508–513, 2014.
doi: 10.1509/jmr.14.0020.

[524] Ji Soo Yi.
Visualized decision making: development and application of information visualization techniques to improve decision quality of nursing home choice.
PhD thesis, Georgia Institute of Technology, 2008.

[525] Ji Soo Yi, Rachel Melton, John Stasko, and Julie A Jacko.
Dust & magnet: multivariate information visualization using a magnet metaphor.
Information Visualization, 4(4):239–256, 2005.

[526] Ji Soo Yi, Youn ah Kang, and John Stasko.
Toward a deeper understanding of the role of interaction in information visualization.
IEEE transactions on visualization and computer graphics, 13(6):1224–1231, 2007.

[527] Robert B Zajonc.
Mere exposure: A gateway to the subliminal.
Current directions in psychological science, 10(6):224–228, 2001.

[528] Dan Zakay and Stuart Wooler.
Time pressure, training and decision effectiveness.
Ergonomics, 27(3):273–284, 1984.

[529] Dimitris Zavras, Vasiliki Tsiantou, Elpida Pavi, Katerina Mylona, and John Kyriopoulos.
Impact of economic crisis and other demographic and socio-economic factors on self-rated health in greece.
The European Journal of Public Health, 23(2):206–210, 2012.

[530] Yunfeng Zhang, Rachel KE Bellamy, and Wendy A Kellogg.
Designing information for remediating cognitive biases in decision-making.
In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2211–2220. ACM, 2015.

[531] Jason Chen Zhao, Wai-Tat Fu, Hanzhe Zhang, Shengdong Zhao, and Henry Duh.
To risk or not to risk?: Improving financial risk taking of older adults by online social information.
In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 95–104. ACM, 2015.

[532] Jian Zhao, Michael Glueck, Simon Breslav, Fanny Chevalier, and Azam Khan.
Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations.
IEEE Transactions on Visualization and Computer Graphics, 23(1):261–270, 2017.

[533] Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, and Baoquan Chen.
Visual clustering in parallel coordinates.
In *Computer Graphics Forum*, volume 27, pages 1047–1054. Wiley Online Library, 2008.

[534] Hartmut Ziegler, Tilo Nietzschmann, and Daniel A Keim.
Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments.
In *Information Visualisation, 2008. IV'08. 12th International Conference*, pages 287–295. IEEE, 2008.

[535] Stanley Zionts.
Mcdm—if not a roman numeral, then what?
Interfaces, 9(4):94–101, 1979.

[536] Torre Zuk and Sheelagh Carpendale.
Visualization of uncertainty and reasoning.
In *International Symposium on Smart Graphics*, pages 164–177. Springer, 2007.

[537] ΠΛΑΤΩΝ.
ΠΛΑΤΩΝ: ΠΟΛΙΤΕΙΑ.
ΠΟΛΙΣ, 2014.

Titre : La visualisation d'information pour la prise de décision: identifier les biais et aller au-delà du paradigme de l'analyse visuelle

Mots clés: visualisation d'information, prise de décision, biais cognitifs, effet d'attraction, interaction homme-machine, économie comportementale

Certains problèmes ne peuvent être résolus ni par les ordinateurs seuls ni par les humains seuls. La visualisation d'information est une solution commune quand il est nécessaire de raisonner sur de grandes quantités de données. Plus une visualisation est efficace, plus il est possible de résoudre des problèmes complexes. Dans la recherche en visualisation d'information, une visualisation est généralement considérée comme efficace quand elle permet de comprendre les données. Les méthodes d'évaluation cherchent à déterminer si les utilisateurs comprennent les données affichées et sont capables d'effectuer des tâches analytiques comme, par exemple, identifier si deux variables sont corrélées.

Cette thèse suggère d'aller au-delà de ce ``paradigme de l'analyse visuelle" et élargir le champ de recherche à un autre type de tâche: la prise de décision. Les tâches de décision sont essentielles à tous, du directeur d'entreprise qui doit prendre des décisions importantes à l'individu ordinaire qui choisit un plan de carrière ou désire simplement acheter un appareil photo. Néanmoins, les décisions ne se résument pas à la simple compréhension de l'information et sont difficiles à étudier. Elles peuvent impliquer des préférences subjectives, n'ont pas toujours de vérité de terrain, et dépendent souvent de connaissances externes aux données visualisées. Pourtant, les tâches de décision ne font pas partie des taxonomies de tâches en visualisation et n'ont pas été bien définies. De plus, la recherche manque de métriques, de méthodes et de travaux empiriques pour valider l'efficacité des visualisations pour la prise de décision.

Cette thèse offre une définition opérationnelle pour une classe particulière de tâches de décision, et présente une analyse systématique qui identifie les visualisations multidimensionnelles compatibles avec ces tâches. Elle présente en outre la première comparaison empirique de techniques de visualisation multidimensionnelle basée sur leur capacité à aider la décision, et esquisse une méthodologie et des métriques pour évaluer la qualité des décisions. Elle explore ensuite le rôle des instructions dans les tâches de décision et des tâches analytiques équivalentes, et identifie des différences de performance entre les deux tâches.

De même que les sciences de la vision informent la visualisation d'information sur les limites de la vision humaine, aller au-delà du paradigme de l'analyse visuelle implique de prendre en compte les limites du raisonnement humain. Cette thèse passe en revue la théorie de la décision afin de mieux comprendre comment les humains prennent des décisions, et formule une nouvelle taxonomie de biais cognitifs basée sur la tâche utilisateur. En outre, elle démontre empiriquement que des biais peuvent être présents même quand l'information est bien visualisée, et qu'une décision peut être ``correcte" mais néanmoins irrationnelle, dans le sens où elle est influencée par des informations non pertinentes.

Cette thèse examine finalement comment mitiger les biais. Les méthodes pour améliorer le raisonnement humain reposent souvent sur un entraînement intensif à des principes et à des procédures abstraits, qui se révèlent souvent peu efficaces. Les visualisations offrent une opportunité dans la mesure où ses concepteurs peuvent remodeler l'environnement pour changer la façon dont les utilisateurs assimilent les données. Cette thèse passe en revue la théorie de la décision pour identifier de possibles solutions de conception. De plus, elle démontre empiriquement que supplémenter une visualisation par des interactions qui facilitent des stratégies de décision alternatives peut mener à des décisions plus rationnelles.

Via des études empiriques, cette thèse suggère que le paradigme de l'analyse visuelle n'est pas en mesure de relever tous les défis de la prise de décision aidée de la visualisation, mais qu'aller au-delà peut contribuer à faire de la visualisation un puissant outil de prise de décision.

Title : Information visualization for decision making: iIdentifying biases and moving beyond the visual analysis paradigm

Keywords: information visualization, decision making, cognitive biases, attraction effect, human-computer interaction, behavioral economics

There are problems neither humans nor computers can solve alone. Computer-supported visualizations are a well-known solution when humans need to reason based on a large amount of data. The more effective a visualization, the more complex the problems that can be solved. In information visualization research, to be considered effective, a visualization typically needs to support data comprehension. Evaluation methods focus on whether users indeed understand the displayed data, can gain insights and are able to perform a set of analytic tasks, e.g., to identify if two variables are correlated.

This dissertation suggests moving beyond this "visual analysis paradigm" by extending research focus to another type of task: decision making. Decision tasks are essential to everybody, from the manager of a company who needs to routinely make risky decisions to an ordinary person who wants to choose a career life path or simply find a camera to buy. Yet decisions do not merely involve information understanding and are difficult to study. Decision tasks can involve subjective preferences, do not always have a clear ground truth, and they often depend on external knowledge which may not be part of the displayed dataset. Nevertheless, decision tasks are neither part of visualization task taxonomies nor formally defined. Moreover, visualization research lacks metrics, methodologies and empirical works that validate the effectiveness of visualizations in supporting a decision.

This dissertation provides an operational definition for a particular class of decision tasks and reports a systematic analysis to investigate the extent to which existing multidimensional visualizations are compatible with such tasks. It further reports on the first empirical comparison of multidimensional visualizations for their ability to support decisions and outlines a methodology and metrics to assess decision accuracy. It further explores the role of instructions in both decision tasks and equivalent analytic tasks, and identifies differences in accuracy between those tasks.

Similarly to vision science that informs visualization researchers and practitioners on the limitations of human vision, moving beyond the visual analysis paradigm would mean acknowledging the limitations of human reasoning. This dissertation reviews decision theory to understand how humans should, could and do make decisions and formulates a new taxonomy of cognitive biases based on the user task where such biases occur. It further empirically shows that cognitive biases can be present even when information is well-visualized, and that a decision can be ``correct" yet irrational, in the sense that people's decisions are influenced by irrelevant information.

This dissertation finally examines how biases can be alleviated. Current methods for improving human reasoning often involve extensive training on abstract principles and procedures that often appear ineffective. Yet visualizations have an ace up their sleeve: visualization designers can re-design the environment to alter the way people process the data. This dissertation revisits decision theory to identify possible design solutions. It further empirically demonstrates that enriching a visualization with interactions that facilitate alternative decision strategies can yield more rational decisions.

Through empirical studies, this dissertation suggests that the visual analysis paradigm cannot fully address the challenges of visualization-supported decision making, but that moving beyond can contribute to making visualization a powerful decision support tool.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France