

Master Thesis

The Credibility of Recommender Systems:
Identifying biases and overspecialisation

Author

Akansel Özgören

Faculty

Behavioural, Management and Social Sciences (BMS)

Programme

MSc in Business Administration

Specialisation

Strategic Marketing & Digital Business

Examination committee

Dr. A.B.J.M. Wijnhoven

Dr. M. de Visser

Date

19th May 2020

Version

Final

Acknowledgements

I would like to thank Dr. A.B.J.M. Wijnhoven for his academical advice and support during my entire graduation period. I also would like to thank Dr. M. de Visser for his additional feedback which has helped me to finalise this master thesis. Additionally, I would like to express my gratitude to my family, friends and all the others who have supported me throughout my years as a student. Without their support, my accomplishments and graduation would not have been possible. I wrote this master thesis with the intention of awakening individuals, providing them with knowledge, and helping them to make relevant decisions.

Akansel Özgören
Deventer, 19th May 2020
akanselozgoren@hotmail.com

Abstract

Recommender systems (RS) are artificial intelligence techniques that aim to reduce information overload and to provide users with diverse, serendipitous, and relevant recommendations in several application domains. However, there are still RS that only operate to increase the income of merchants without inspiring users to make relevant decisions. These RS provide users with biased and overspecialised recommendations which can lead to manipulation, irrelevant decisions, and low customer satisfaction. The motive of this study is to create a mechanism that allows users to identify biases and overspecialisation within RS so that they can avoid these potential problems and make relevant decisions. Based on the message credibility and triangulation theory, a bias & overspecialisation identification tool (BOIT) has been developed and used within an online experiment with 82 participants. The findings of this experiment indicate that participants were able to identify types of bias and overspecialisation within an e-commerce recommender system. As a result, the credibility of this recommender system decreased significantly. Therefore, it is concluded that the BOIT spreads awareness among users about potential biases and overspecialisation within RS and that it has a statistically significant effect on users' judgment of the credibility of RS.

Keywords: Recommender Systems, Artificial Intelligence, E-Commerce, Bias, Overspecialisation, Message Credibility, Triangulation.

Table of contents

Acknowledgements	2
Abstract	3
1. Introduction	7
2. Problem analysis.....	9
2.1 Filtering algorithms	9
2.2 Weaknesses of the filtering algorithms	10
2.3 Biased recommendations.....	12
2.4 Overspecialisation	13
2.5 Summary of the problem analysis	14
3. Theory	17
3.1 Message credibility and triangulation	17
3.2 Theory classification	19
3.3 Amazon and Tweaklers	21
3.4 Hypotheses and conceptual model	22
4. Methodology.....	24
4.1 Research design.....	24
4.2 Selection and sample	24
4.3 Operationalisation and measurement	24
4.4 Data collection and analysis	26
5. Results	28
5.1 Pre-test.....	28
5.2 BOIT usage.....	28
5.3 Post-test	31
5.4 Reliability	32
5.5 Validity.....	32
6. Discussion & conclusion.....	34
6.1 Key findings	34
6.2 Limitations and future research	35
6.3 Implications	36
Reference list.....	37
Appendices	42
Appendix I: Paper selection procedure for chapter 2: Problem analysis.....	42
Appendix II: Sponsored recommendations	43
Appendix III: Rating types within RS	44
Appendix IV: Demographic data of the sample	44
Appendix V: Usage of BOIT.....	46

Appendix VI: LR χ^2 3x3 contingency tables	48
Appendix VII: Inter-item correlations credibility scale	52
Appendix VIII: Questionnaire	53

List of figures

Figure 1. Conceptual model.	23
Figure 2. BOIT formative indicators.	26
Figure 3. Three-item credibility scale.	26
Figure 4. Effect size equation (Field, 2009).	27
Figure 5. BOIT ‘yes-scores’.	29
Figure 6. BOIT ‘maybe-scores’.	29
Figure 7. BOIT ‘no-scores’.	30

List of tables

Table 1. Filtering algorithms.	10
Table 2. Weaknesses of the filtering algorithms.	11
Table 3. Problems of RS with possible solutions.	15
Table 4. Key concepts of the problem analysis.	16
Table 5. Formative indicators of message credibility (Appelman & Sundar, 2016).	17
Table 6. Reflective indicators of message credibility (Appelman & Sundar, 2016).	18
Table 7. Types of triangulation (Wijnhoven & Brinkhuis, 2015).	18
Table 8. Data triangulator.	19
Table 9. Theory triangulator.	19
Table 10. Investigator triangulator.	20
Table 11. Methods triangulator.	20
Table 12. Relevance triangulator.	21
Table 13. Alignment of reflective indicators and triangulators.	21
Table 14. BOIT checklist.	25
Table 15. Pre-test results Amazon.	28
Table 16. Pre-test results Tweakers.	28
Table 17. BOIT totals Amazon.	30
Table 18. BOIT totals Tweakers.	30
Table 19. LR χ^2 values Amazon.	31
Table 20. LR χ^2 values Tweakers.	31
Table 21. Post-test results Amazon.	31
Table 22. Post-test results Tweakers.	31
Table 23. Post-test changes.	32
Table 24. Summary hypothesis tests.	32
Table 25. Results of Cronbach’s α	32

List of abbreviations:

RS:	Recommender Systems
E(-commerce):	Electronic
AI:	Artificial Intelligence
MAUT:	Multiple Attribute Utility Technique
HyPER:	Hybrid Probabilistic Extensible Recommender
BOIT:	Bias & Overspecialisation Identification Tool
H(1):	Hypothesis
N:	Sample Size
LR:	Likelihood Ratio
df:	degrees of freedom
M:	Mean
SD:	Standard Deviation

1. Introduction

The numbers of electronic commerce (e-commerce) organisations have been increasing since the development of the World Wide Web (WWW). The Internet, as a marketing channel, is different in comparison with the traditional retail channels (Park, Lee, & Han, 2006). Consumers that regularly shop online, cannot touch, or smell the products. Due to this, they need to base their judgments only on the information about the product, which is presented on the websites of the e-commerce organisations. The enormous growth of this available information, which is also powered by the rapid adoption of the internet, is making access to relevant information more difficult than before. This phenomenon caused the information overload problem (Arazy, Kumar, & Shapira, 2010; O'Donovan & Smyth, 2005). Recommender systems (RS) are artificial intelligence (AI) techniques that are used as tools to interact with large and complex information spaces and to minimise information overload by helping consumers to access products and services that suit their requirements ideally (Burke, Felfernig, & Göker, 2011; Montaner, López, & De La Rosa, 2003; Teppan & Zanker, 2015). Within this study, the term 'RS' will be used to abbreviate recommender systems. RS are key components of successful online shops (Arazy et al., 2010). According to Aggarwal (2016), the primary goal of RS is increasing product sales of merchants. Besides this, RS also have operational and technical goals. Aggarwal (2016) states that RS aim to deliver recommendations that are relevant, serendipitous and diverse for users. Lu, Wu, Mao, Wang and Zhang (2015) state that RS are mainly used in the following eight domains: E-government, e-business, e-commerce, e-library, e-learning, e-tourism, e-resource services and e-group activities. Moreover, it is indicated that recommendations from RS have a notable influence on consumer's preferences, willingness to pay and their choices (Adomavicius, Bockstedt, Curley, & Zhang, 2019; Milano, Taddeo, & Floridi, 2019).

There are different types of RS. Those different types will be elaborated in further detail within the upcoming chapters of this study. Besides, the contemporary weaknesses of RS will be discussed as well. The main weaknesses of RS are the following types of bias, which are still ubiquitous within RS: Rating bias, serial position effects, decoy effects, risk aversion and popularity bias (Abdollahpour, Burke, & Mobasher, 2017; Adomavicius et al., 2019; Teppan & Zanker, 2015). In addition, overspecialisation is also still ubiquitous within RS, which results in low user satisfaction (Adamopoulos & Tuzhilin, 2015; Kotkov, Wang, & Veijalainen, 2016). As a consequence, the presence of the biases and overspecialisation within RS allows third-party agents to manipulate their recommender system to make sure that it operates in their favour (Adomavicius et al., 2019). This phenomenon results in a loss of credibility in the RS and it harms the long-term value that it can deliver to users if users find out that the recommendations are biased.

The motive of this study is to decrease the effect of manipulation of RS by spreading awareness among users about the types of bias and overspecialisation within RS. To accomplish this, a bias & overspecialisation identification tool (BOIT) will be created and applied by users. The BOIT will be created in a way that it is ready to be applied in several application domains and that it is understandable and easy to apply. After this, users are allowed to judge the credibility of RS more easily since they are able to identify biases and overspecialisation. Additionally, after judging the credibility of RS, users can decide if they want to neutralise them. In other words, users can choose to neutralise RS simply by ignoring them and by making use of other more credible RS. Finally, to test what the effects are of the BOIT on the judgment of the credibility of RS, the following central research question of this study will be answered.

“What are the effects of the BOIT on users' judgment of the credibility of recommender systems?”

This study aims to provide the academic field of business administration, e-business, and information systems with crucial information regarding the ubiquitous biases and overspecialisation within RS and how these can be identified by users. Besides this, this study aims to deliver new academic insights by developing a mechanism based on the classification of RS credibility theories. Furthermore, this study aims to have societal relevance by spreading awareness among users of RS about biased and overspecialised recommendations to decrease manipulation and irrelevant decisions.

This master thesis is structured as follows. The second chapter consists of the problem analysis. The types of RS (filtering algorithms), their weaknesses, biased recommendations, and overspecialisation within RS will be elaborated and discussed within the problem analysis. Within the theory chapter, the used RS credibility theories will be clarified. Next, the hypotheses, conceptual model, and the two RS that will be used to test the hypotheses will be presented. Within the methodology chapter, the research design, sample data, data collection and data analysis will be presented. Within the results chapter, the results of the experiment will be reported, and the hypotheses will be tested. Subsequently, the reliability and validity of the experiment will be assessed. The final chapter will consist of the key findings, limitations, ideas for future research and the implications of this study.

2. Problem analysis

Within this chapter, the problems that led to the formation of the central research questions will be elaborated. RS are distinguished as filtering algorithms. Within the first two sections, the different types of filtering algorithms will be explained, and their weaknesses and potential solutions will be provided in detail. Next, the types of bias and overspecialisation within RS will be presented and elaborated. This chapter will end with a summary of the problem and a list of key concepts. To select the most suitable papers for the problem analysis, the guidelines of Kitchenham and Charters (2007) were used (Appendix I).

2.1 Filtering algorithms

To develop functioning RS, a few steps need to be followed. The first step is the profile representation, which creates the user profile (Montaner et al., 2003). Moreover, RS need to gather information from users, such as interests, to provide them with the wanted results from the beginning. Due to this, RS need to make use of a suitable technique that will help them generate an accurate initial profile for users. Burke and Ramezani (2011) argue that RS need to have social knowledge about the larger community of users and RS need to have individual knowledge about target users. To collect this information, RS can gather relevance feedback to learn the interests of users. Mostly, the feedback which is offered implicitly or explicitly by the user has no sense (Montaner et al., 2003). Therefore, a profile learning technique is needed. This profile learning technique extracts and structures the relevant information depending on the representation of the user's profile. If the interests of users will change, the user profile needs to change as well to retain the desired accuracy in its exploitation and a technique that adapts the user profile to the new interests (Montaner et al., 2003). After developing the user profile, they will be exploited, and the RS will provide recommendations to users that consist of items. The word 'item' is the term that is used to signify what the system recommends to users, such as products or services (Ricci, Kantor, Rokach, & Shapira, 2011).

To recommend items to users, different types of filtering algorithms are applied by RS. The three main information filtering algorithms of RS are demographic filtering, content-based filtering and collaborative filtering. (Bobadilla, Ortega, Hernando, & Alcalá, 2011; Montaner et al., 2003; Pazzani, 1999). The demographic filtering algorithm applies descriptions of users of the RS to learn the relationship between items and the types of users who will probably like them (Montaner et al., 2003). This approach is established on the assumption that individuals with common attributes such as gender, age and nationality will have the same common preferences. In other words, this filtering algorithm creates user profiles through stereotypes. RS also need content knowledge about the recommended items (Burke & Ramezani, 2011). The content-based filtering algorithm provides users with recommendations by analysing the description of the items that have been rated by the target user and the description of the items to be recommended (Montaner et al., 2003). User profile-item matching methods can be used to compare the interests of the users with the right items. Moreover, content-based filtering recommended items are similar to the items that the target user liked in the past (Bobadilla et al., 2011; Huang, 2011; Ricci et al., 2011). The most commonly used and studied filtering algorithm within RS is collaborative filtering (Bobadilla et al., 2011). The collaborative filtering algorithm creates recommendations by finding correlations among other users of the RS. This approach uses feedback from a set of people concerning a set of items to make recommendations (Montaner et al., 2003). This means that collaborative filtering is the process of filtering items by using the opinions of other people (Schafer, Frankowski, Herlocker, & Sen, 2007). Ekstrand, Riedl and Konstan (2011) describe different types of collaborative filtering in their paper. The user-user collaborative filtering algorithm finds other users with a rating history close to that of the target user and ultimately uses their ratings on other items to predict items that the target user will like. On the contrary, item-item collaborative filtering uses similar ranking patterns of items. In addition, Ekstrand et al. (2011) state it is expected that users have similarities among their preferences for comparable items.

Burke (2002) and Huang (2011) discuss the utility-based filtering algorithm. Utility-based RS create recommendations that are focused on the calculation of the utility of each item for users. This approach applies the user profile as the utility function that the system has derived from users. The Multiple Attribute Utility Technique (MAUT) is often used as a technique to generate utility-based

recommendations. The MAUT takes various attributes and objectives that might have a high level of utility for users by analysing the strengths and weaknesses of these attributes and objectives (Sudesh, Dharmic, Pulari, & Ramesh, 2018). Moreover, it can also factor non-product attributes into the utility calculations such as product availability and vendor reliability. Knowledge-based filtering is the fifth filtering algorithm that will be described here. Burke (2002) and Ricci et al. (2011) state that this filtering algorithm is similar to the utility-based approach since it also aims to recommend items that could meet the need of users. Additionally, this approach also has no issues with new users and items. However, the knowledge-based approach is distinguished in that it has functional knowledge (Burke, 2002). This means that this approach knows how a particular item could meet a particular need of a user. Namely, it explains the relationship between a need and a potential recommendation (Burke, 2002). Finally, community-based filtering, also called social network-based filtering, is the last filtering algorithm that will be described here. Community-based RS recommend items based on the rating preferences of the social network of the target user (Arazy et al., 2010; Fatemi & Tokarchuk, 2013; Lu et al., 2015). Community-based RS can be compared to collaborative RS since they both combine users. However, community-based RS are more trust-based because they combine users with their social media friends, instead of combining them with users that they do not know personally (Lu et al., 2015). All the described filtering algorithms are summarised in Table 1.

Filtering algorithms	Provides users with recommendations by...
Demographic	...establishing the assumption that individuals with common attributes such as gender, age and nationality will have the same common preferences.
Content-based	... analysing the description of the items that have been rated by the user and the description of the items to be recommended.
Collaborative	...using input from a collection of people on a set of items to find correlations among other users of RS.
Utility-based	...calculating the utility of each item for users based on the user profile and the MAUT.
Knowledge-based	...calculating the utility of each item for users based on functional knowledge.
Community-based	... recommending items based on the ratings and preferences of their social network.

Table 1. Filtering algorithms.

As told in the introduction chapter, RS are used in a broad variety of application domains. Lu et al. (2015) state that RS with filtering algorithms such as collaborative, content-based, and knowledge-based still play a dominant role in nearly all application domains. Besides, they state that RS in the e-learning domain have highly applied knowledge-based methods, whereas e-resource RS have more collaborative based methods. According to Montaner et al. (2003), e-commerce RS are based on history-based profile representation models. Thus, those RS barely use any profile learning techniques. Therefore, Montaner et al. (2003) state in their paper that most of the e-commerce RS make use of content-based filtering. Nowadays, this statement is not relevant anymore since e-commerce sites made major efforts to understand the user better by employing new profile learning techniques to provide users with more appropriate recommendations (Singh & Mehrotra, 2016).

2.2 Weaknesses of the filtering algorithms

Within the previous section, six types of filtering algorithm were discussed. However, the filtering algorithms are not perfect and do have their weaknesses. Demographic filtering can lead to an incorrect representation of the world due to a large amount of generalisation (Montaner et al., 2003). Besides, the demographics do not change together with their interests, but they remain static over time. With content-based filtering, subjective characteristics are not considered because of the objective content. Additionally, there is a lack of ‘randomness’. This means that this approach recommends more of what the user has already observed and indicated as a preferred item (Montaner et al., 2003). This could eventually lead to a massive filter bubble. Furthermore, Montaner et al. (2003) state that the recommender quality of the content-based filtering approach is not frequently accurate if there is a low number of item ratings. The collaborative filtering approach is considered as the most used filtering

algorithm according to the scientific literature. However, it also has its disadvantages. Collaborative filtering cannot accurately find similar users for target users with unique interests, which results in non-accurate recommendations (Montaner et al., 2003). In addition, collaborative filtering has the early-rater and few-user problem. The early-rater problem refers to items that cannot be recommended because they are not rated. The few-user problem refers to items that cannot be recommended properly if there is a low number of users. Those two problems are also known as the cold-start problem (Madadipouya & Chelliah, 2017). Besides the cold-start problem, collaborative filtering RS also suffer from data sparsity. Data sparsity refers to the complexity of finding a sufficient and reliable number of similar users, as users regularly rate a small part of the items (Guo, Zhang, & Thalmann, 2014). The utility-based filtering algorithm does not have issues with cold-start and sparsity because the recommendations are not based on accumulated statistical evidence (Burke, 2002). However, users need to build a complete preference function and weigh each attribute's importance by him or herself (Huang, 2011). Therefore, it requires an enormous amount of human interaction which is also expensive (Sudesh et al., 2018). Knowledge-based RS are generally designed for domains with highly customised items, which makes it difficult for rating information to directly reflect greater preferences (Aggarwal, 2016). In community-based RS, the recommendations depend on the social network of users. Victor, Cornelis and De Cock (2011) indicate that cold-start users in collaborative RS are often also cold-start users in the context of community-based RS. They claim that new users need to be encouraged to connect to other users so they can expand their network as soon as possible. Additionally, Ahmadian et al. (2020) state that recommendations of community-based RS are heavily depended on the availability of social networks. They argue that users who have expressed many social relationships are likely to have many ratings. The weaknesses of the filtering algorithms are summarised in Table 2.

Filtering algorithms	Weaknesses
Demographic	Large generalisation and static demographics.
Content-based	Subjective characteristics are not considered, lack of randomness and lack of preciseness of recommender quality.
Collaborative	Non-accurate recommendations for users with unique interests, cold-start problem, and data sparsity.
Utility-based	Without (expensive) human interaction, the utility of an item cannot be calculated.
Knowledge-based	Difficult for rating information to directly reflect greater preferences in highly customised domains.
Community-based	Cold-start problem and heavily depended on the availability of social networks.

Table 2. Weaknesses of the filtering algorithms.

To solve the weaknesses of each filtering algorithm, Adomavicius and Tuzhilin (2005), Burke (2002), Çano and Morisio (2017), Montaner et al. (2003) and Ricci et al. (2011) propose to combine two or more filtering algorithms to create hybrid RS. Ricci et al. (2011) provide an example of a hybrid recommender system where a collaborative and content-based approach were combined to solve the following problems: The collaborative filtering approach suffers from the cold-start problem and can therefore not recommend items without ratings. However, this does not restrict the content-based filtering approach because of the estimation of new items is based on their features which are generally easily accessible. Hybrid RS are typically designed for specific problem domains. Nevertheless, they can be limited in their ability to generalise to other settings and therefore cannot frequently make use of further information. For this reason, Kouki, Fakhraei, Foulds, Eirinaki, and Getoor (2015) developed a general-purpose, extensible system that makes use of arbitrary data modalities aiming to enhance the recommendations provided to users. They propose a general hybrid recommender system called HyPER, which stands for Hybrid Probabilistic Extensible Recommender. It combines multiple different sources of information and modelling techniques into one model. Kouki et al. (2015) set up their system by applying probabilistic soft logic, which is an intuitive probabilistic programming language. Applying probabilistic soft logic enables efficient and accurate predictions. Therefore, they claim that it can outperform existing filtering algorithms.

2.3 Biased recommendations

In some circumstances, RS may also be a source of manipulation. Adomavicius, Bockstedt, Curley, Zhang, and Ransbotham (2019) claim that RS do more than just reflect user preferences. Instead, they shape them. RS have the potential to courage biases and, for example, affect sales of e-commerce organisations in unexpected ways. As a result, RS can manipulate preferences in ways users do not recognise. Adomavicius et al. (2019) state that online recommendations significantly affect the willingness to pay when users know less about items. This allows unethical organisations to manipulate their recommendations to gain more profit. In another study by Adomavicius et al. (2019), it is claimed that the word ‘bias’ is considered disapproving and representative of a negative prejudice. Furthermore, they claim that RS could be biased if users only receive high and unprofessional system-predicted ratings. Besides this, users seem to rate items higher that already have a high rating (Adomavicius et al., 2019). This can distort or manipulate the preferences and the item choices of users in a way that potentially will lead to irrelevant decisions. As told in the introduction chapter, this could reduce the level of credibility of the RS if users know that these recommendations are biased. Besides this, it may harm the long-term value that it can deliver to users.

Next to rating bias, there are four more types of bias within RS: Serial position effects, decoy effects, risk aversion and popularity bias (Abdollahpouri et al., 2017; Teppan & Zanker, 2015). Teppan and Zanker (2015) discuss the first three types in their paper. Serial position effects describe the phenomenon that items at the beginning (primacy) and at the end (recency) of the list are more likely to be remembered by users than those in the middle (Felfernig et al., 2007). This can be the case if certain items are sponsored by the source of the recommender system and therefore placed at the beginning of a recommendation list. An example of this is presented in Appendix II. Decoy effects increase the attraction of predefined items. On the other hand, they decrease the attraction of the items of competitors and the list of recommended items will be less complete due to the exclusion of those competitive items. In RS with decoy effects, the strengths of the predefined items are compared with the weaknesses of competing items. Thus, there is an unfair comparison. If the decoy effects of RS are strong, users will not have the possibility to rate the utility of the items in an objective way. This may lead to poor decisions (Teppan & Felfernig, 2012). Moreover, Teppan and Zanker (2015) argue that users tend to experience losses more than gains. This initiates users to react risk-averse at moments when items are labelled in terms of gains and risk-seeking. Since users tend losses more than gains, they will eventually choose for the less risky item, even if the expected level of utility is lower than the riskier option. This an example of risk aversion, which is also called ‘framing’. Popularity bias is discussed in the paper of Abdollahpouri et al. (2017). They claim that collaborative filtering algorithms often emphasise popular items, that have more ratings, over other less popular items, the so-called long-tail items. Those long-tail items, for example, niche items, are only popular by a small group of users. The popular items are also likely well-known products. Because of this, there is a lack of novelty and the recommendations may have a low level of serendipity. In addition, the RS will ignore the interests of users that are attracted to niche items.

Overall, most of the biases within RS need to be identified by users themselves. Milano et al. (2019) state that the influence that RS have on users deserves ethical scrutiny. The potential biases need to be understood and addressed by users. In the paper of Kaptein, Markopoulos, De Ruyter and Aarts (2015), it is argued that organisations could also use RS as personalised persuasive systems that use persuasion profiles. The authors provide an example in their paper of a system that applied short persuasive messages for users to reduce their unhealthy snacking behaviour. It can be said that this way of influencing is more ethical since the system encourages users to live healthier lives. Nevertheless, Kaptein et al. (2015) also argue that there are still uncertainties regarding ethics and privacy that need to be addressed if designers of persuasive systems want to apply personalised persuasion.

With the purpose to reduce biases, Adomavicius et al. (2019) distinguish different types of ratings: numerical, graphical, star and binary (Appendix III). There is evidence that graphical rating display designs of RS are more beneficial than numerical designs in reducing biases in RS. Adomavicius et al. (2019) state these designs led to lower biases in the post-consumption preference ratings of users. However, none of the types of ratings can remove biases completely. Moreover, Teppan and Zanker (2015) argue that there is strong domination of RS risk aversion strategies. In addition, serial position effects are the most recessive out of the three types of bias. Besides, serial position and decoy effects are only relevant when risk aversion is not prevalent. Finally, traditional RS do not have the technical

capabilities to control these three types of bias and that these three types of bias are ubiquitous in RS. Therefore, Teppan and Zanker (2015) note that it is necessary to provide users with a mechanism that allows the identification and neutralisation of disingenuous biases to enable users to make more objective decisions when they interact with RS. By doing this, the persuasive power of RS can be released. Teppan and Felfernig (2012) present an approach that neutralises decoy effects. This decoy minimisation approach restores objectivity by removing items from the item set or by adding decoys such that the influences dominate each other. Further, Abdollahpouri, Burke and Mobasher (2019) demonstrate a post-processing step that manages popularity bias and can be utilised in the output of RS. It enables RS to accomplish the desired trade-off between accuracy and better coverage of the less popular products that are stuck in the long tail of item popularity. Abdollahpouri et al. (2019) note that their approach focusses on recommending long-tail items while keeping the loss of accuracy small compared to traditional RS.

2.4 Overspecialisation

The low level of unexpectedness and serendipity of certain recommendations that leads to low user satisfaction levels is defined as overspecialisation (Kotkov et al., 2016). Adamopoulos and Tuzhilin (2015) note that various RS provide users with items that are already familiar with the items that the user has bought. Due to this, there is a low interest to these items and the recommendations will not have a large impact on the behaviour of users. Adamopoulos and Tuzhilin (2015) provide the following example in their paper: RS may recommend products such as milk and bread to users. Despite the fact of being precise, in the sense that the users will indeed buy these two products, such recommendations are of little interest since they are conspicuous, because the users will, most likely, buy these products even without these recommendations. Adamopoulos & Tuzhilin (2015) claim in their paper that the notion of unexpectedness is a key dimension of improvement that significantly contributes to the overall performance and usefulness of RS. Overspecialised RS also lack serendipity. Serendipitous recommendations involve novel items with a low discovery probability (Adamopoulos & Tuzhilin, 2015). De Gemmis, Lops, Semeraro and Musto (2015) identify serendipity as recommendations that try to help users to find items that are interesting for them and that they might not have discovered by themselves. Besides, Maksai, Garcin and Faltings (2015) identify serendipity as both unexpected and useful. Moreover, De Gemmis et al. (2015) provide the following example where they demonstrate a recommender system with an overspecialisation problem that fails to provide users with serendipitous recommendations: RS with collaborative filtering algorithms will search for similar products that a user has liked by suggesting products by other people who liked the same product. Because of the similarity, the recommended product will be likely a known product to the user which will result in a low level of serendipity.

If users frequently receive expected and non-serendipitous recommendations, they can end up in a filter bubble. Kamishima, Akaho, Asoh, and Sakuma (2012) define a filter bubble as a selection of the appropriate diversity of information provided to users. Lately, the provided information to users is becoming restricted to the information that is initially preferred by them. This restriction occurs due to the influence of personalised technologies. Therefore, users will be placed in a separate bubble (Pariser, 2011). Because of the restriction of these bubbles, users will lose the opportunity of finding new items. Zuiderveen Borgesius et al. (2016) provide an example with a personalised news website. This website may prioritise liberal or conservative media items, depending on the presumed political interests of the users. As a consequence, users may receive a small selection of political items from only one specific point of view, rather than more or even all points of view. Furthermore, users prefer to receive content they feel familiar with and viewpoints that they agree with (Nagulendra & Vassileva, 2014). However, this leads to the existence of filter bubbles where users will be filtered away and they will live in echo chambers where they are exposed to conforming opinions (Flaxman, Goel, & Rao, 2016).

To decrease overspecialisation, RS aim to provide users with a diverse range of unexpected and serendipitous recommendations. Badran, Bou abdo, Al Jurdi and Demerjian (2019) claim that higher user satisfaction can be realised by including serendipity at the cost of profile accuracy. To realise this, the expectations of the users need to be clear. Zhou, Xu, Sun and Wang (2017) propose a serendipitous new recommendation algorithm. The proposed model is based on a collaborative filtering approach and follows three aspects: Unexpectedness, insight, and value of an item. ‘Insights’ stand for the importance

of the ability to relate a new clue to experience and knowledge in the occurrence of serendipity. ‘Value’ demonstrates the relation between the value of the provided information and the potential needs and concerns of users. Badran et al. (2019) apply different aspects in their algorithm for serendipitous recommendations. They vary the serendipity and accuracy ratio to achieve the ideal number of serendipitous recommendations. This algorithm has three steps: Quality calculations, unexpectedness calculation, and utility calculation. With the quality calculation, a lower quality limit for the recommended items is fixed. The item’s quality is compared with the lower limit. The item continues to the next step if its quality is higher. With the unexpectedness step, the expected recommendations will be calculated. Then, the range of unexpectedness will be calculated. If the items belong to the range of unexpectedness, they continue to the last step. The last step, utility calculation, estimates the utility of the items for users. Items with the highest utility will be recommended to provide users with serendipitous and unexpected items.

Looking at the filter bubble, Nagulendra and Vassileva (2014) aim to decrease filter bubbles with interactive visualisation. The design and implementation of the visualisation of the filter bubbles are based on personalised stream filtering, which is an implementation of a privacy-aware decentralised social network that uses an open-source framework. Furthermore, Bozdag and van den Hoven (2015) investigate tools that aim to decrease filter bubbles. They state that most of the tools do not disclose their objectives and do not specifically describe the filter bubble. Bozdag and van den Hoven specifically studied the weaknesses of the tools. As an example, they claim that the visualisation tool of Nagulendra and Vassileva (2014) does not try to support users into challenging information. With this tool, users can decide to remain in the filter bubble. As told earlier, filter bubbles provide users with items that are already familiar to users. Therefore, it can be said that the serendipity of the recommendations is low. Matt, Benlian, Hess and Weiß (2014) state that filter bubbles can be decreased by serendipitous recommendations. This leads to a higher level of perceived fit and enjoyment. In addition, de Gemmis et al. (2015) state that the determination of the filter bubble and the process of finding unexpected recommendations out of the bubble is one of the most common strategies of the programming process of serendipitous RS.

2.5 Summary of the problem analysis

The filtering algorithms, their weaknesses, overspecialisation, and the types of bias within RS are now all presented and discussed within the problem analysis. This chapter will summarise the discussed main problems within RS. The discussed problems of RS and their possible solutions are reported in Table 3. A list of the key concepts of the problem analysis is presented in Table 4.

After discussing the creation of the user profile, the following types of filtering algorithms were presented and discussed together with their weaknesses: Demographic, content-based, collaborative, utility-based, knowledge-based, and community-based. Designers of RS can fix the weaknesses of the filtering algorithms by combining several filtering algorithms to create hybrid RS. Ricci et al. (2012) provided an example with a hybrid recommender system that combined the content-based filtering algorithm with the collaborative filtering algorithm to solve the weaknesses of both filtering algorithms. Kouki et al. (2015) propose in their paper a general-purpose, extensible framework for hybrid RS which they call HyPER. The results of their study reveal that this approach outperforms standard hybrid RS on efficiency and accuracy.

Bias and overspecialisation are both still ubiquitous within RS. Overspecialisation can be reduced by providing users with unexpected and serendipitous recommendations. This can be achieved by understanding serendipity and the expectations of the users, to avoid that they end up in filter bubbles and echo chambers. Looking at biased RS, it can be argued that there is no single solution that can entirely fix this problem yet. Adomavicius et al. (2019) demonstrated that RS with graphical rating display design could decrease the level of bias. However, this design is not able to remove biases completely. Adomavicius et al. (2019) also claimed that biases could allow third-party agents to manipulate the RS to make sure that it operates in their favour. This could lead to a loss of credibility in the RS. Another type of bias, which is based on the popularity of items, could be decreased by boosting items that are less popular to deliver serendipitous recommendations to the user (Abdollahpouri et al. 2019). Kaptein et al. (2015) proposed a persuasive system that can influence users more ethically. Nevertheless, there are still uncertainties regarding ethics and privacy that need to be addressed if designers of persuasive systems want to apply personalised persuasion. Finally, Milano et al. (2019)

state that users need to scrutinise RS on ethics, and Teppan and Zanker (2015) claim in their paper that users need a mechanism to identify and neutralise potential biases in RS to lower the persuasive power of RS so that they would not misinterpret the recommendations. Due to this, the BOIT will be developed and tested in the upcoming chapters of this study.

Problems of RS	Possible solutions
Weaknesses of the filtering algorithms	Hybrid filtering algorithms and HyPER.
Biased recommendations: rating, serial position, decoy, risk aversion and popularity	Identification and neutralisation.
Overspecialisation: lack of unexpected and serendipitous recommendations. As a result of this, users end up in filter bubbles and echo chambers	Identification, neutralisation, gathering data about the expectations of the users and calculating quality, unexpectedness, and utility of the item.

Table 3. Problems of RS with possible solutions.

Concept	Definition
Recommender systems (RS)	RS are AI techniques that are used as tools to interact with large and complex information spaces and to ease information overload by helping consumers to find products and services that suit their requirements ideally (Burke, Felfernig, & Göker, 2011; Montaner, López, & De La Rosa, 2003; Teppan & Zanker, 2015).
Users	Ricci et al. (2011) define ‘users’ as the individuals that use RS. Users have diverse goals and characteristics. To personalise the recommendations, RS exploit information about different users (Montaner et al., 2003).
Items	The word ‘item’ is the term that is used to signify what the system recommends to users, such as products or services (Ricci et al., 2011).
Rating bias	Adomavicius et al. (2019) claim that RS could be biased if users only receive high and unprofessional system-predicted ratings. Besides this, users seem to rate items higher that already have a high rating (Adomavicius et al., 2019). This can distort or manipulate the preferences and purchases of users in a way that potentially will lead to poor item choices.
Serial position effects	Serial position effects refer to the phenomenon that items at the beginning (primacy) and at the end (recency) of the list are more likely to be remembered by users than those in the middle (Felfernig et al., 2007). RS can use serial position effects to present predefined items in the beginning or at the end of a recommendation list to persuade users to buy these items.
Decoy effects	Decoy effects increase the attraction of predefined items. On the other hand, they decrease the attraction of the items of competitors and the list of recommended items will be less complete due to the exclusion of those competitive items. In RS with decoy effects, the strengths of the predefined items are compared with the weaknesses of competing items. Thus, there is an unfair comparison. If the decoy effects of RS are strong, users cannot rate the utility of the items in an objective way. This may lead to irrelevant decisions (Teppan & Felfernig, 2012).
Risk aversion	Risk-averse RS initiate users to react risk-averse at moments when items are labelled in terms of gains and risk-seeking (Teppan & Zanker, 2015). When users tend losses more than gains, they will eventually choose for the less risky item, even if the expected level of utility is lower than the riskier option.
Popularity bias	RS with popularity bias emphasise popular items with a higher rating over other less popular items, the so-called long-tail items. Those long-tail items are only popular by a small group of users, such as niche items. The popular items are also likely well-known products (Abdollahpouri et al., 2017).
Overspecialisation	Overspecialised RS have a low level of unexpectedness and serendipity. De Gemmis et al. (2015) and Kotkov et al. (2016) define this concept as a

	recommendation that provides users with items within the existing range of their interests. If users regularly receive recommendations that are not unexpected and serendipitous, they will be less satisfied, and they end up in filter bubbles and echo chambers.
Filter bubbles	Kamishima, Akaho, Asoh, and Sakuma (2012) define a filter bubble as a selection of the appropriate diversity of information provided to users. Lately, the provided information to users is becoming restricted to the information that is initially preferred by them. This restriction occurs due to the influence of personalised technologies. Therefore, users will be placed in a separate bubble (Pariser, 2011). Eventually, users will live in echo chambers where they are exposed to conforming opinions (Flaxman et al., 2016).
Identification and neutralisation	Bias and overspecialisation are still ubiquitous within RS (Abdollahpouri et al., 2017; Adomavicius et al., 2019; de Gemmis et al., 2015; Kotkov et al., 2016; Teppan & Zanker, 2015). Therefore, users need to identify the biases and overspecialisation within RS to avoid manipulation and irrelevant decisions. When users identify the biases by using the BOIT, they can decide to neutralise the recommender system. In other words, users can decide to make the biased recommender system ‘harmless’ by not relying on it or even not making use of it. Hence, they can release the persuasive power of RS.

Table 4. Key concepts of the problem analysis.

3. Theory

This chapter will clarify which RS credibility theories will be used and how they will be classified to create the BOIT. Next, the two RS that will be used in the experiment will be presented. Lastly, the hypotheses and conceptual model of this study will be provided.

3.1 Message credibility and triangulation

The BOIT will serve as an understandable, concise, and easy-to-use mechanism that alerts users and allows them to identify biases and overspecialisation so that they can judge RS on credibility. To realise the creation of the BOIT, two RS credibility theories will be used: Message credibility and triangulation. The formative indicators of the message credibility theory will serve as a set of quality requirements of bias-free, unexpected, and serendipitous RS. After applying the BOIT, the credibility of the RS will be judged by applying the three-item credibility scale with reflective indicators of the message credibility theory. The second theory that will be applied is the triangulation theory. The triangulation theory refers to the combination of several research methodologies and their application in the study of the same phenomenon (Denzin, 2015). Wijnhoven and Brinkhuis (2015) distinguish five types of triangulators: data, theory, investigator, method, and relevance. Moreover, the formative indicators of message credibility will be divided into the set of triangulators. Thus, every triangulator can be applied so that all types of bias and overspecialisation that were discussed within the problem analysis can be identified. The classification of the formative indicators will be based on the definition of the formative indicators in the context of RS, and the requirements of each triangulator. After the classification, the types of bias and overspecialisation will be aligned with the suitable formative indicators. In the upcoming paragraphs, the two theories will be explained in more detail.

Appelman and Sundar (2016) define message credibility as: “The individual’s judgment of the veracity of the content of communication” (p. 63). They present a scale with quality requirements of the credibility of news articles in their paper. This scale is parsimonious, reliable, valid, and useful in multiple situations where manipulated messages could appear. The quality requirements of the message credibility theory are divided into two groups: The formative and reflective indicators. Appelman and Sundar (2016) state that formative indicators include objective measures of quality, expertise, and fairness of a message. On the other hand, reflective indicators are indicators that determine the level of credibility of a message. The formative and reflective indicators are presented below in Table 5 and 6. The results of the study of Appelman and Sundar (2016) reveal that message credibility can be measured with a study by asking participants to rate how well the indicators describe the received content. Therefore, it can be said that this scale is a useful measure for different studies of message credibility.

Formative indicators	Definition
Complete	These indicators contribute to perceptions of credibility as a sense of fairness.
Concise	
Consistent	
Well-presented	
Objective	These indicators underscore the need for impartiality on the part of the RS.
No spin	
Representative	This indicator suggests the importance of achieving balanced coverage by representing multiple sides of a problem.
Expert	These two indicators factor into user conceptions of message credibility.
Will have impact	
Professional	Professionalism is a significant predictor of message credibility.

Table 5. Formative indicators of message credibility (Appelman & Sundar, 2016).

Reflective indicators	Definition
Accuracy	These three indicators describe the content and reflect the concept of message credibility and make all three sense of the proposed definition of message credibility. ‘Accuracy’ and ‘authenticity’ could be seen as more objective. On the other hand, ‘believability’ could be seen as more subjective. However, the three-item credibility scale is based on self-report perceptions. Thus, it can be said that the three indicators are all subjective.
Authenticity	
Believability	

Table 6. Reflective indicators of message credibility (Appelman & Sundar, 2016).

Triangulation is a method that enhances the reliability of the results of a study and enables to saturate the data (Fusch, Fusch, & Ness, 2018). In addition, Wijnhoven and Brinkhuis (2015) argue that the use of triangulation leads to better insights of the real world. Besides, they argue that it leads to better decisions, gaining a more complete and integrated perspective on phenomenon’s and more consciously developing opinions on topics. Denzin (as cited in Fusch et al., 2018; Wijnhoven & Brinkhuis, 2015) has developed four types of triangulation that can be used to improve the objectivity, credibility and validity of data. The different triangulators are described in Table 7. Additionally, Wijnhoven and Brinkhuis (2015) found, based on the inquiring systems, that there is also a fifth triangulator: Relevance.

Triangulator	Definition	Inquiring system requirements
Data	A representativeness check of obtained data and the quality and precision of observation, the constancy over numerous observations, and the non-appearance of theoretical and normative bias.	Lockean: Verify data validity, check data reliability, and precision.
Theory	Identification of basic assumptions and norms, the inclusion and exclusion of variables, and the relations among variables. Besides this, theory triangulation identifies the perspectives of the published document.	Leibnizian: Identify variables, causalities, goals, and values. Kantian: Identify perspective, ontology, and categories.
Investigator	Focuses on the knowledge about the interests of the author or publisher from which biases can be uncovered. To receive more diversity of opinions on a topic, authors and publishers with opposing interests and positions need to be found.	Hegelian: Identify author, publisher, expertise of author, site reputation, author’s affiliation(s), the interests of an author, an author’s sentiment and presenting opposing views.
Methods	Identification of scope, grounding theory, ontology, used categories, research method and replications.	Kantian: Identify the research method and document replications.
Relevance	This triangulator is related to the others since it requires input from them to decide on the usefulness of internet information.	Singerian: testing the usefulness of internet information, the effectiveness of the solution, is open to multiple perspectives, innovative, adaptive, and ideal in complex situations.

Table 7. Types of triangulation (Wijnhoven & Brinkhuis, 2015).

Furthermore, Wijnhoven and Brinkhuis (2015) state that inquiring systems provide requirements for the types of triangulators and information quality. Inquiring systems describe the ideas of five influential western philosophers (Locke, Leibniz, Kant, Hegel, and Singer) from the perspective of systems theory (Churchman, 1971; Courtney, 2001; Mason & Mitroff, 1973; Wood, 1983). Each inquiring system provides a solution for a different problem by starting with different primitive elements or building blocks (Mason & Mitroff, 1973). Inquiring systems are used as theoretical support for the dimensions of triangulation because they propose teleological systems for the creation of knowledge that also includes norms for information quality (Wijnhoven & Brinkhuis, 2015). Therefore, each triangulator received requirements that are based on inquiring systems.

Data triangulation is based on the Lockean inquiring system since they both check the validity, reliability, and precision of the data. Investigator triangulation is based on the Hegelian inquiry system because they both specifically investigate the author or publisher. Methods triangulation is based on the Kantian inquiring system since they both identify the relevant categories of ontology to allow the individual to evaluate the wholeness of a perspective in a document. Theory triangulation is based on the Leibnizian and Kantian inquiring systems since theory triangulation is in line with the requirements of these two systems. The relevance triangulator is based on the Singerian inquiring system because they both need effective use from the feedback from the other triangulators and inquiring systems to make decisions.

3.2 Theory classification

Within this section, the formative indicators of message credibility will be classified into the triangulators. Next, the types of bias and overspecialisation will be aligned with the formative indicators so that the indicators will contribute to the identification of biases and overspecialisation. The five triangulators will be presented separately in the tables below. Finally, the three reflective indicators of credibility scale will be specifically defined and aligned with the triangulators.

Decoy effects and popularity bias are aligned with the formative indicator ‘complete’. The RS need to be checked if it presents a complete list of recommended items and not with predefined or popular items with increased attraction, such as sponsored items. The complete list needs to contain both popular and less popular (niche) items to decrease popularity bias. Decoy effects are also aligned with ‘representative’. The representativeness check of the obtained items is important since users will then be exposed to a higher range of different items, which results in balanced coverage (Appelman & Sundar, 2016). This indicator is aligned with decoy effects since RS with decoy effects refuse to recommend competing items that represent the main item well in terms of content (Teppan & Zanker, 2015).

Data triangulator (Lockean)	
Formative indicator	Bias(es)
Complete	Decoy effects and popularity bias
Representative	Decoy effects

Table 8. Data triangulator.

Serial position effects are aligned with the formative indicator ‘consistent’. The recommended items have to be presented to users consistently. This means that the order of the items needs to be randomised every time a user interacts with it, to avoid serial position effects. Risk aversion is aligned with ‘concise’. If RS apply a risk aversion strategy, additional messages are added (Teppan & Zanker, 2015). Because of this, RS become less concise. A message needs to be concise because then it will contribute to perceptions of message credibility (Appelman & Sundar, 2016). Rating bias is aligned with the formative indicator ‘well-presented’. The recommended items need to be well-presented with graphical display designs to decrease bias since this is claimed in by Adomavicius et al. (2019).

Theory triangulator (Leibnizian and Kantian)	
Formative indicator	Bias(es)
Consistent	Serial position effects
Concise	Risk aversion
Well-presented	Rating bias

Table 9. Theory triangulator.

Decoy effects are aligned with the formative indicator ‘objective’. This formative indicator underscores the impartiality on the part of the designer of RS. RS need to be objective by presenting items that have no increased attraction by added decoys. As stated in the problem analysis, RS that only recommend items of one brand, are also called decoy effects. Risk aversion is aligned with ‘no-spin’. No-spin also underscores the impartiality on the part of the investigator, so on the part of the designer of the RS (Appelman & Sundar, 2016). If RS apply a risk aversion strategy, they need to be impartial by removing the risk-averse items. Teppan and Zanker (2015) argued in their paper that users mostly act

risk-aware and tend to go for the less risky option, even if the level of utility is high. Therefore, credible RS need to act spin-free and need to be honest about the real level of risk and utility of the item for the user. The utility of the item for the user can be calculated by applying the MAUT (Sudesh et al., 2018). Popularity bias is aligned with the indicator ‘expert’. If RS can also boost and recommend long-tail items to users to reduce popularity bias, they will have a high expert-level. By doing this, more diversity among recommended items will be created and opposing, less popular (niche) items will be added into the recommendation list.

Investigator triangulator (Hegelian)	
Formative indicator	Bias(es)
Objective	Decoy effects
No-spin	Risk aversion
Expert	Popularity bias

Table 10. Investigator triangulator.

Rating bias is aligned with the formative indicator ‘professional’. Identifying the research method is a task that needs to be done according to the Kantian inquiring system, which is linked to the methods triangulator (Wijnhoven & Brinkhuis, 2015). With this triangulator, users can identify the research method of the user ratings for the items of RS. Users need to check the professionalism of the ratings of the items. RS are professional if the ratings are based on appreciative inquiry, interviews, focus groups, case studies, action research, questionnaires, surveys, experiments, observational studies, secondary data, literature studies, sampling, or structural equations. This list of research methods is indicated in the paper of Wijnhoven and Brinkhuis (2015).

Methods triangulator (Kantian)	
Formative indicator	Bias(es)
Professional	Rating bias

Table 11. Methods triangulator.

The last triangulator will be aligned with the concepts of overspecialisation. Unexpectedness and serendipity are aligned with the formative indicator ‘will have impact’. To avoid overspecialisation, the recommended items need to be unexpected or serendipitous for users. If this is the case, the recommended items will have impact on the behaviour of users. Specifically, the recommendations will be useful since the findings from the problem analysis revealed that unexpected and serendipitous recommendations are more useful than overspecialised recommendations. For example, in the paper of Adamopoulos & Tuzhilin (2015), it is claimed that the notion of unexpectedness is a key dimension of improvement that significantly contributes to the overall performance and usefulness of RS. Besides this, Maksai et al. (2015) identified the concept of serendipity as unexpected and useful. Filter bubbles and echo chambers are aligned with ‘fair’. This indicator is not presented in Table 5. However, it is still relevant since it can be aligned with filter bubbles and echo chambers. RS need to be exposed to a different and fair range of items so that users will not only receive the same recommended items and that they will not be exposed to conforming opinions about these items. The final indicator that will be added to the relevance triangulator, is ‘ease of use’. This indicator is not discussed by Appelman and Sundar and not aligned with one certain type of bias or overspecialisation. However, after using the other triangulators, the user can decide if the recommender system was easy to use. The ‘ease of use’ indicator will be a part of the relevance triangulator because it requires the effective use of all the other triangulation methods. Furthermore, ease of use is also related to usefulness, which is related to the formative indicator ‘will have impact’. According to Davis (1989), perceived usefulness and perceived ease of use are essential elements of user acceptance of information technology. Finally, it is expected that the indicators that represent one type of bias are associated since they measure the same concept. The indicators ‘complete’ and ‘objective’ measure decoy effects. Therefore, it is expected that they are associated with each other. In other words, it is expected that users who think that a certain recommender system is not complete also think that the same recommender system is not objective.

Relevance triangulator (Singerian)	
Formative indicator	Overspecialisation
Will have impact	Unexpectedness and serendipity
Fair	Filter bubbles and echo chambers
Ease of use	The effective use of all the triangulators

Table 12. Relevance triangulator.

The five triangulators are now all aligned with formative indicators and the types of biases and overspecialisation. Now, the three reflective indicators of the credibility scale will be clarified. The reflective indicator ‘accuracy’ will be used to evaluate RS if they meet the technical and operational goals of RS, which were presented by Aggarwal (2016). If recommendations are relevant, serendipitous, and diverse for users, then the recommendations meet the technical and operational goals and are therefore accurate. The accuracy of RS can be measured more easily after applying the relevance triangulator since this triangulator checks the serendipity (‘will have impact’) and diversity (‘fairness’) of RS. To check if RS are authentic, the data, theory, investigator, and methods triangulator need to be applied to check if biases are ubiquitous within RS. Fewer types of bias result in higher authenticity of RS and a higher level of trust since biases decrease the level of the trust of users in RS (Adomavicius et al., 2019). The last reflective indicator, the believability of RS, is entirely subjective since believability is considered more subjective than the other two indicators (Appelman & Sundar, 2016). Users base the judgment of this believability of RS purely on their perceptions.

Reflective indicator	Definition
Accuracy	Recommendations that are relevant, serendipitous, and diverse for users (relevance triangulator).
Authenticity	Recommendations that are bias-free and are therefore trusted (data, theory, investigator, and methods triangulator).
Believability	The subjective opinion of users about the believability of RS: Do users believe that RS recommend items that are truly useful for them?

Table 13. Alignment of reflective indicators and triangulators.

3.3 Amazon and Tweakars

Within this study, the recommender system of Amazon and Tweakars will be used and rated by participants in an experiment. Both RS are used within the e-commerce domain. Amazon was founded in 1994 by Jeff Bezos (DePillis & Sherman, 2019) and is currently the largest e-commerce organisation in the world (Bhasin, 2019). In its early years, Amazon started its business by selling books, TV shows, and films via their website. Nowadays, Amazon broadened its services by increasing their range of products. The organisation now also sells electronics, food, grocery, clothing, and many more under one roof. In 1998, Amazon launched its first recommender system (Smith & Linden, 2017). This recommender system is still being used today and is based on the item-item collaborative filtering algorithm and ‘star only’ item ratings. The recommender system uses the purchase history of users, browsing history, the current item users are viewing, and the behaviour of other users as data to recommend items (Simran, Pande, & Desai, 2019). The recommender system of Tweakars was the other recommender system in the experiment. This Dutch organisation was founded in 1998 by Femme Taken (Taken, 2008). Tweakars is a technology review website that aims to provide consumers with information about hardware, software and the internet by testing and reviewing the latest products. Additionally, Tweakars is an independent organisation. This means that another organisation cannot buy a positive review score, organisations cannot offer money to test their products, and the price comparison tool on the website stays independent as well (Tweakars, 2020a). The recommender system of Tweakars displays the item ratings in the ‘star only’ style and makes use of the hybrid filtering algorithm. Their website mentions that the recommender system recommends popular items (collaborative filtering algorithm) with similar prices and specifications to the main item (content-based filtering algorithm) (Tweakars, 2020b). The two RS are illustrated in Appendix VIII.

Those two RS were chosen for the experiment because the two organisations and the two RS differ from each other. Amazon is a commercial organisation that aims to make a profit by selling its products, while Tweakars is independent and wants to provide users only with information about items.

Based on the problem analysis and the BOIT, it can already be said that the recommender system of Amazon has more potential biases and overspecialisation than that of Tweakers. Firstly, it can be said that the recommender system of Amazon is less complete than the one of Tweakers since it only recommends highly rated items from one brand: Samsung. The recommender system of Tweakers also recommends items from other brands such as LG, Phillips, and Sony. Due to this, the recommender system of Amazon lacks completeness, objectivity, expertise, and representativeness and, thus, has potential decoy effects and popularity bias. Amazon's recommender system also recommends items that are not in line with the main item since it also recommends items of the same brand with other measurements than 55-inch, while Tweakers only recommends 55-inch alternatives. Based on this, it can be said that the recommender system of Amazon is unfair since it only recommends items from one brand, even items that are not similar to the main item in terms of specifications. Therefore, this supports the idea that users can end up in a filter bubble while using the recommender system of Amazon. Nevertheless, Tweakers' recommender system is not perfect. Based on the theory of serial position effects, it is noteworthy to state that the first four items that are listed in the recommendation list of Tweakers are all items from the brand LG. This supports the idea that the recommender system of Tweakers lacks consistency and has potential serial position effects. The study will reveal if users will identify the same biases and overspecialisation within the two RS that are mentioned here above.

3.4 Hypotheses and conceptual model

As told in section 3.2, it is expected that the indicators that are aligned with one type of bias are associated and therefore dependent on each other since they measure the same concept. Moreover, based on the provided information in section 3.3, the recommender system of Amazon reveals more types of bias and overspecialisation within their recommendations than Tweakers. Thus, it is also expected that the recommender system of Amazon will score lower on credibility after applying the BOIT in the post-test. On the other hand, it is expected that the credibility score of the recommender system of Tweakers will increase after applying the BOIT in the post-test since it has fewer types of bias and overspecialisation within their provided recommendations. The hypotheses of this study are presented here below. Figure 1 illustrates the conceptual model of this study. It indicates that the usage of the BOIT affects the perceived accuracy, authenticity, believability, and therefore the credibility of the recommender system since the three reflective indicators are the determinants of credibility (Appelman & Sundar, 2016).

H1: “The formative indicators, which are aligned together with one type of bias, are significantly associated with each other.”

H1a: “‘Well-presented’ and ‘professional’ are significantly associated with each other since they are both aligned with rating bias.”

H1b: “‘Complete’ and ‘objective’ are significantly associated with each other since they are both aligned with decoy effects.”

H1c: “‘Complete’ and ‘representative’ are significantly associated with each other since they are both aligned with decoy effects.”

H1d: “‘Objective’ and ‘representative’ are significantly associated with each other since they are both aligned with decoy effects.”

H1e: “‘No-spin’ and ‘concise’ are significantly associated with each other since they are both aligned with risk aversion.”

H1f: “‘Complete’ and ‘expert’ are significantly associated with each other since they are both aligned with popularity bias.”

H2: “The BOIT significantly affects users’ judgment of the credibility of the RS.”

H2a: “The credibility of the recommender system of Amazon will significantly decrease after applying the BOIT in the post-test.”

H2b: “The credibility of the recommender system of Tweakers will significantly increase after applying the BOIT in the post-test.”

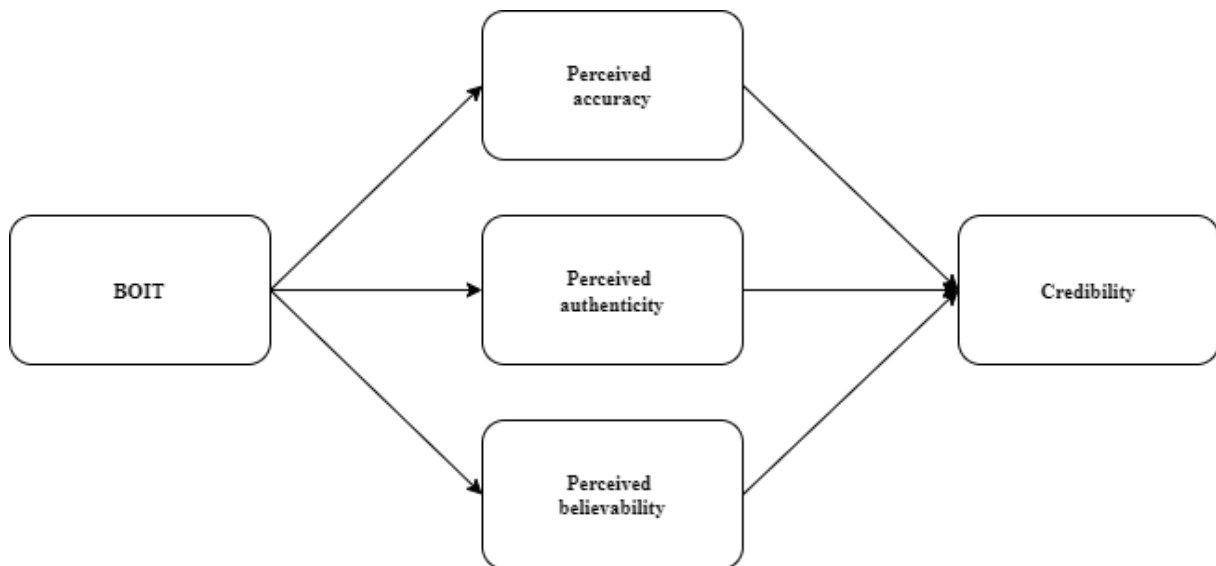


Figure 1. Conceptual model.

4. Methodology

Within the methodology chapter, the research design for this study will be explained. Next, the sample data will be provided. After this, operationalisation and measurement will be discussed. Subsequently, the BOIT and the three-item credibility scale that the participants have applied during the experiment will be illustrated. Finally, the applied statistical tests, reliability test, and validity assessments will be explained within the data collection and analysis section.

4.1 Research design

To test the hypotheses, an online within-subjects experiment was conducted. In a within-subject experiment, each individual is exposed to more than one of the treatments being tested (Charness, Gneezy, & Kuhn, 2012). The choice for a within-subjects experiment was made due to its main advantages, compared to a between-subjects experiment research design, where the behaviour of participants in one experimental condition is compared to those in another. Charness et al. (2012) state that a within-subjects design offers a substantial boost in statistical power, internal validity does not depend on random assignment, and that there will be an exact comparison since every participant will see and rate the same treatments. Therefore, the required sample size compared to a between-subjects design is also lower (Charness et al., 2012). Within this experiment, participants were exposed to a scenario where one main item was chosen. Namely, the participants were told that they were searching for a 55-inch television. The participants received two recommendations lists, one of Amazon's recommender system and one of Tweakers' recommender system. The main item that they received was a Samsung 55-inch television and the list with the recommended items below the main item was investigated by the participants. The experiment consisted of a pre- and post-test. In the pre-test, the participants were asked to judge the RS only with the three-item credibility scale, so without the BOIT. In the post-test, the BOIT was presented to the participants and they were asked to identify potential biases and overspecialisation within both RS before judging the same RS again on credibility.

4.2 Selection and sample

Brysbaert (2019) argues that a minimum number of 52 participants is needed for a within-subjects design with two levels to reach an effect size of $d = .4$, which is a small to medium effect (Cohen, 1992). Therefore, the desired minimum sample size of this within-subjects experiment was $N = 52$. The participants were randomly recruited through Facebook, LinkedIn, and Instagram. Besides this, participants were also recruited through the BMS subject-tool, SONA. The distribution of the experiment started on 5th March 2020 and ended on 25th March 2020. The experiment was completed by 82 participants, which resulted in a sample size of $N = 82$. Thus, the objective to have a minimum sample size of $N = 52$ was completed.

The sample was almost evenly split in terms of gender: 39 males and 43 females. The ages ranged from 18 to 55, but approximately 4 out of 5 participants were younger than 26 ($M = 25.67$, $SD = 9.72$). Most of the participants follow or have finished a study on bachelor's level ($n = 63$). This is mainly the case because 47 of the 82 participants were recruited through SONA, which is commonly used by bachelor students from the University of Twente. The other participants follow or have finished a study on associate level (Dutch: MBO) ($n = 12$), master's level ($n = 6$) and less than high school level ($n = 1$). Most of the participants were Dutch ($n = 38$) or German ($n = 36$). Other participants were Turkish ($n = 4$), Italian ($n = 2$), Latvian ($n = 1$) or Lithuanian ($n = 1$). A more detailed overview of the demographic data of the sample is presented in Appendix IV.

4.3 Operationalisation and measurement

The terms 'user' and 'item' were replaced with 'consumer' and 'product' just for the experiment to make these terms more understandable for the participants. In the pre- and post-test, participants applied the three-item credibility scale, which consists of the three reflective indicators of message credibility (Appelman & Sundar, 2016). The three reflective indicators of the credibility scale were briefly defined to the participants and the rating scale, which was also applied in the study of Appelman and Sundar (2016), was applied by the participants to judge the credibility of RS: 1: very poorly to 7: very well.

This scale provides a parsimonious and usable metric for determining the credibility of data for use in academic and industry research. Moreover, according to Appelman and Sundar (2016), the sum of the scores of the three reflective indicators determine the credibility score. Therefore, participants could assign a credibility score of 3 to 21 for each recommender system. In the post-test, the BOIT (Table 14) was provided to the participants. In the BOIT, the formative indicators are aligned with the types of bias and overspecialisation and they are presented in bold italics. Participants were asked the following question for each indicator: “Does this formative indicator contribute to the credibility of the recommender system?” They had the following answer options: ‘Yes, ‘maybe’ or ‘no’. After applying the BOIT, the participants were asked again to rate the credibility of both RS again. The first four participants were asked to provide comments on the clarity of the questions and the BOIT before answering the questions in the post-test. The first participant mentioned before the start of the post-test that the BOIT consisted of long and repetitive sentences. Subsequently, the sentences were shortened so that the BOIT was easier to read for the participants while answering the questions in the post-test. Moreover, all the 82 participants applied the same BOIT (Table 14) during the experiment and no comments were received referring to difficulties with applying the BOIT or with answering the questions. The formative indicators with the answer options and the three-item credibility scale are illustrated in Figure 2 and 3. The complete questionnaire of the experiment was available in English and Dutch and is presented in Appendix VIII.

Bias and overspecialisation	Definition
Rating bias	<i>Poorly presented</i> with numerical ratings and <i>unprofessional</i> (not based on studies or research).
Serial position effects	<i>Inconsistent</i> list with predetermined products in the beginning and the end of the list.
Decoy effects	<i>Incomplete</i> list with only predetermined products without competing products. This is not <i>objective</i> and not <i>representative</i>
Risk aversion	Less risky products with low utility. This is a type of <i>spin manipulation</i> and makes RS less <i>concise</i> .
Popularity bias	<i>Incomplete</i> list with only popular and well-known products. <i>Expert</i> RS include also less-known products from different brands.
Unexpectedness	Unexpected and unknown (serendipitous) products that are useful <i>will have an impact</i> on the behaviour of consumers.
Filter bubble and echo chamber	When consumers receive an <i>unfair</i> selection of products from only one or two brands with strengthened persuasion.

Table 14. BOIT checklist.

	Yes	Maybe	No
Complete	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Representative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Consistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Concise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Well-presented	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Objective	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No spin manipulation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Professional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Will have impact	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ease of use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2. BOIT formative indicators.

	1	2	3	4	5	6	7
Accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Authenticity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Believability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 3. Three-item credibility scale.

4.4 Data collection and analysis

The experiment was conducted by using the software of the online survey tool Qualtrics. The software of Qualtrics is a worldwide leading survey tool of functionalities, security and privacy measures and (BMS, 2020). The experiment was anonymous, and the participants were not asked to use their names during the experiment due to data minimisation and privacy. The collected data were analysed with the software of IBM SPSS Statistics 25. First, a chi-square test was used to determine if the indicators from H1 are significantly associated with each other. Two assumptions need to be met before using the chi-square test. The first assumption is the independence of data. This assumption has been met since each participant contributes to only one cell of the contingency tables. To meet the second assumption, only 20% of the expected frequencies are allowed to be below an expected value of 5. This assumption was violated in seven of the twelve used contingency tables. Therefore, the likelihood ratio (LR) χ^2 statistic was applied to determine the level of statistical significance. This is the alternative of the standard chi-square value and is applied in smaller samples where the second assumption has not been met (Field,

2009; Mchugh, 2013). The strength of the associations was calculated by Cramer's V. A Cramer's V value larger than .25 indicates a very strong association between two variables (Akoglu, 2018).

Furthermore, a one-tailed paired samples t-test was used as the statistical test to test H2. SPSS only provides the two-tailed p-value, but the one-tailed p-value can be obtained by dividing the two-tailed p-value by 2 (Field, 2009). Two assumptions need to be met before conducting a paired samples t-test: The data needs to be measured at an interval or ratio level and the sampling distribution of the differences between scores should be normal (Field, 2009). Looking at the first assumption, all the data for the paired samples t-test are measured at the interval scale because the rating scale is comparable with Likert scale data. Likert scale data are primarily analysed at the interval measurements because they are created by calculating a composite score from four or more types of Likert-type items (Boone & Boone, 2012). This is indeed the case within this experiment. Looking at the second assumption, it can be said that the sampling distribution of the differences between scores is normal since the sample size is considered as large (higher than 50, which is applied as rules of thumb). Therefore, the sampling distribution is approximately normal according to the Central Limit Theorem (Field, 2009; Lumley, Diehr, Emerson, & Chen, 2002). Thus, both assumptions have been met and the paired t-test was allowed to be used. The means (M), standard deviations (SD), t-value, degrees of freedom (df) and the p-value with $\alpha = .05$, will be presented in the tables of the measurements of each hypothesis within the next chapter. Finally, the effect sizes of the tests were also calculated. An effect size of $r = .10$ has a small effect, $r = .30$ a medium effect and $r = .50$ a large effect (Cohen, 1992). The effect sizes were calculated by using the equation in Figure 4.

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Figure 4. Effect size equation (Field, 2009).

Reliability in the research context means that the measure or questionnaire should consistently reflect the construct that it measures (Field, 2009). To assess the reliability within this study, the three-item credibility scale (Figure 3) was assessed with Cronbach's α . If the Cronbach's α is $> .7$ it is acceptable, but it is preferred to have a Cronbach's α which is $> .8$ (Field, 2009). Moreover, a three-item scale with Cronbach's $\alpha = .8$ has an average inter-item correlation of .57 (Cortina, 1993).

The internal, external, and measurement validity of the study will be assessed as well. Internal and external validity are defined as follows. Internal validity refers to the degree to which a study creates a cause-and-effect relationship between the treatment and the observed results (Slack & Draugalis Jr, 2001). An experimental design with high internal validity will deliver replicable and robust results (Schram, 2005). External validity refers to the possibility of generalizing the results to situations that contributed to the study (Schram, 2005). According to Steckler and Mcleeroy (2008), internal validity is broadly considered as the priority of research. However, they also state that the external validity of a study should not be overlooked. They note that if a treatment is effective, it is important to know if the treatment is likely to be effective in other populations and settings. This study focuses mainly on internal validity since the hypotheses refer to the testing of the effectiveness of a treatment, the BOIT. However, external validity will still be assessed to verify if the treatment has the potential to be effective in other settings and other populations. Adcock & Collier (2001) state in their paper that measurement validity is particularly concerned with whether the operationalisation and scoring of cases adequately reflect the concepts that are measured. Measurement validity consists of three different types: construct, criterion, and content validity. Content validity assesses if the content covers all the aspects of the measured concept (Adcock & Collier, 2001). Criterion validity assesses how closely produced scores from an indicator correspond to scores of other variables which are direct measures of the phenomenon of concern (Adcock & Collier, 2001). Finally, construct validity is defined as the degree to which an operationalisation measures the concept that it is expected to measure (Bagozzi, Yi, & Phillips, 1991). The internal, external and measurement validity of the study will be assessed in-depth within the next chapter.

5. Results

Within the first section of this chapter, the credibility scores from the pre-test will be presented. Next, the results of the BOIT usage will be reported and illustrated in bar charts. Subsequently, the likelihood ratio χ^2 scores of the indicators from H1 will be reported. After this, the credibility scores from the post-test will be presented and H2 will be tested. Finally, the reliability and validity of the experiment will be assessed.

5.1 Pre-test

The results of the pre-test of the experiment are reported in Table 15 and 16. The results reveal that Amazon's recommender system was more credible ($M = 13.91$, $SD = 3.422$) compared to Tweakers' recommender system ($M = 13.22$, $SD = 3.975$) in the pre-test.

Amazon	M	SD
Perceived accuracy	4.68	1.266
Perceived authenticity	4.80	1.232
Perceived believability	4.43	1.315
Credibility	13.91	3.422

Table 15. Pre-test results Amazon.

Tweakers	M	SD
Perceived accuracy	4.56	1.389
Perceived authenticity	4.43	1.370
Perceived believability	4.23	1.485
Credibility	13.22	3.975

Table 16. Pre-test results Tweakers.

5.2 BOIT usage

The results of the BOIT usage are presented in Figures 5, 6, and 7. The first Figure presents the numbers of participants who have indicated that the formative indicators contribute to the credibility of both RS. As told in the methodology chapter, participants were asked if the indicators contribute to the credibility of the RS and had three answer options for each indicator: 'Yes', 'maybe' and 'no'. Therefore, each indicator has a maximum frequency of 82 for each recommender system. The figures illustrate that 32 Participants answered that 'complete' contributes to the credibility of the recommender system of Amazon. Additionally, 27 participants answered that 'complete' may contribute to it and 23 participants answered that 'complete' does not contribute to the credibility of Amazon's recommender system ($32+27+23=82$).

The 'fair' indicator reveals the most opposing results in Figure 5. 20 of the 82 participants answered that this indicator contributes to the credibility of Amazon's recommender system, while 45 of the 82 participants answered that this indicator contributes to the credibility of Tweakers' recommender system. Figure 6 presents the numbers of participants who indicated that the formative indicators may contribute to the credibility of both RS. In this Figure, the numbers were all close to each other. The indicators 'consistent' and 'professional' reveal the most contrasting results. Furthermore, Figure 7 reports the numbers of participants who indicated that the formative indicators do not contribute to the credibility of both RS. Figure 7 revealed the most differences between the results. 30 of the 82 participants answered that the indicator 'fair' does not contribute to the credibility of the recommender system of Amazon, while only 6 of the 82 participants answered that this indicator does not contribute to the credibility of the recommender system of Tweakers. Table 17 and 18 reveal the totals of the BOIT usage results per recommender system. There are substantial variations between the total number of participants who answered that the indicators contributed to the credibility of both RS. Tweakers' recommender system received a total of 99 more 'yes-scores' than the one of Amazon. Besides, Amazons' recommender system received a total of 128 more 'no-scores' than that of Tweakers. This corresponds to section 3.4 where it was stated that the recommender system of Amazon has more

potential types of bias and overspecialisation within its recommendations than Tweakers. A more detailed overview of the results of the BOIT usage is presented in Appendix V.

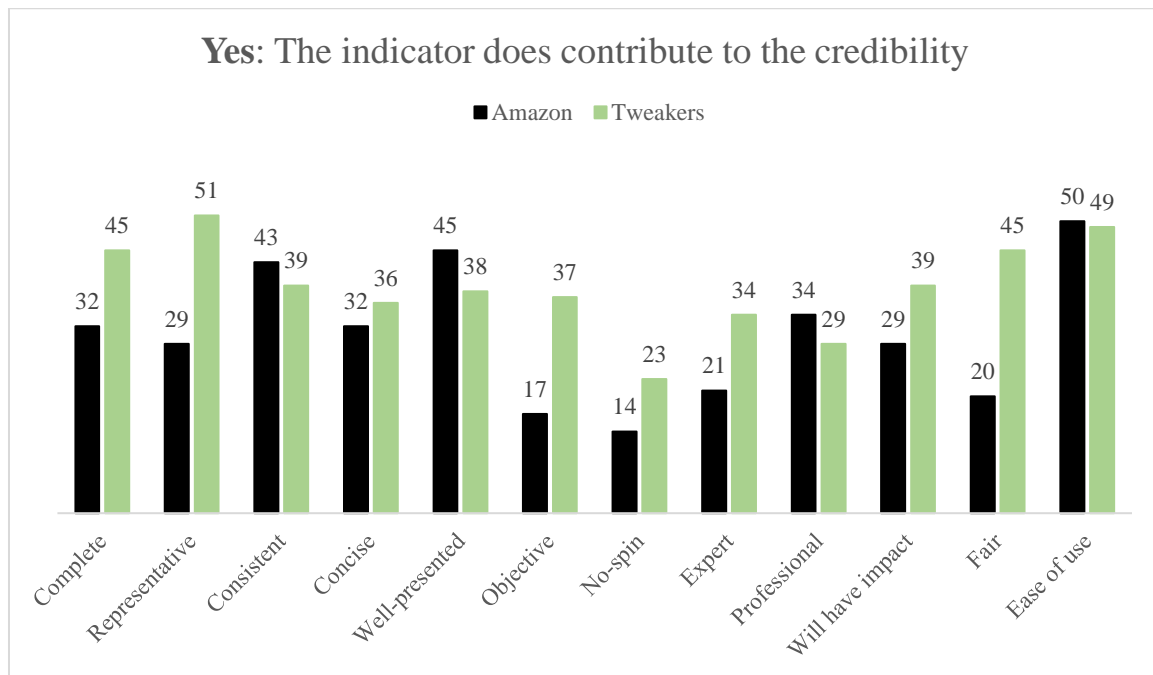


Figure 5. BOIT 'yes-scores'.

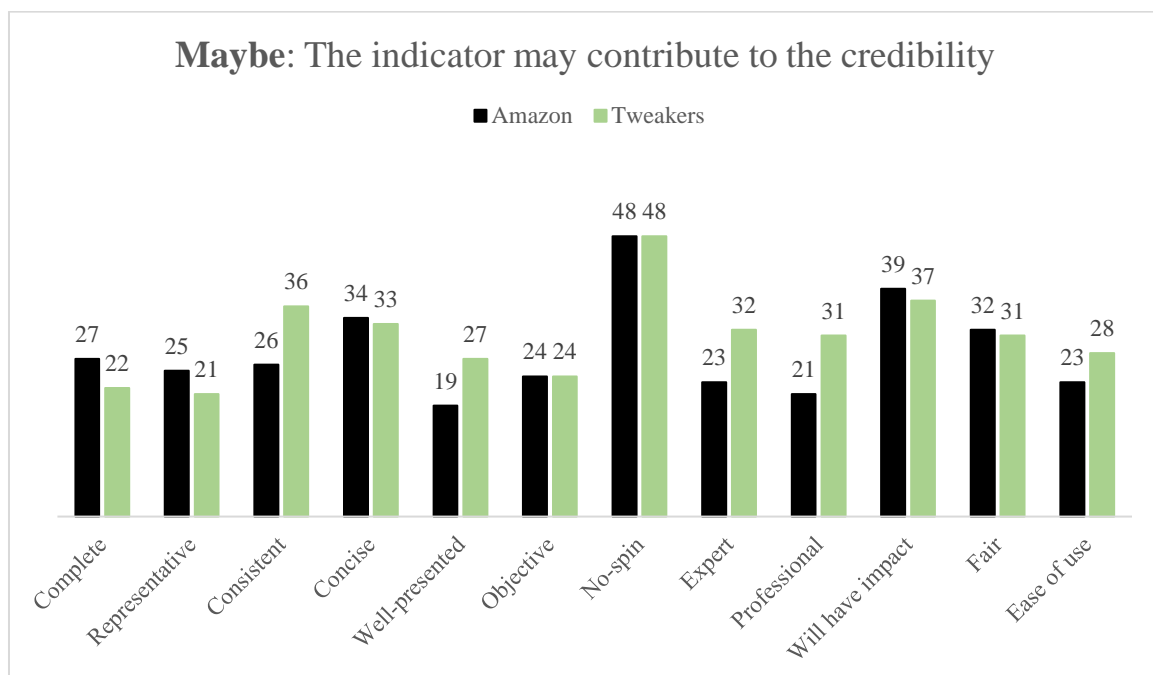


Figure 6. BOIT 'maybe-scores'.

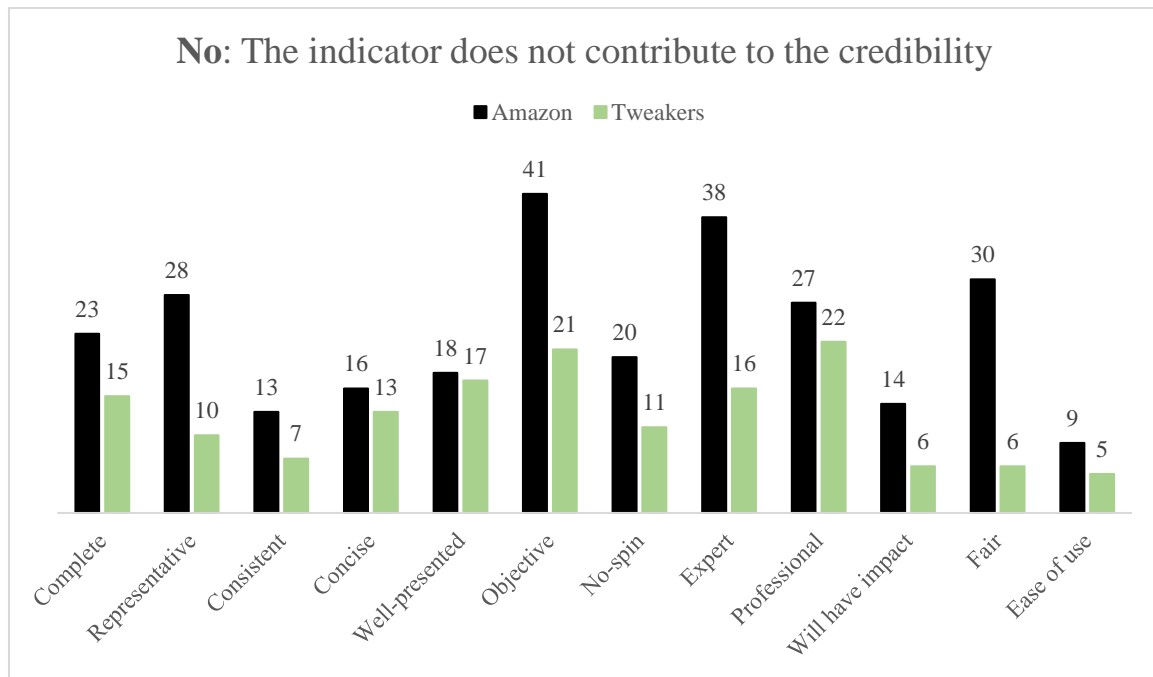


Figure 7. BOIT 'no-scores'.

Amazon	Yes	Maybe	No	Total
Total	366	341	277	984
Per cent	37.20	34.65	28.15	100

Table 17. BOIT totals Amazon.

Tweakers	Yes	Maybe	No	Total
Total	465	370	149	984
Per cent	47.26	37.60	15.14	100

Table 18. BOIT totals Tweakers.

Table 19 reports the LR χ^2 values and the Cramer's V of the aligned indicators that were applied to identify biases and overspecialisation within Amazon's recommender system. Four of the six tests indicate a statistically significant association between the variables. Besides this, the Cramer's V values of those four statistically significant values are all above .25, which means that the associations are very strong (Akoglu, 2018). Table 20 reports the values of Tweakers' recommender system. Five of the six tests indicate a statistically significant association between the variables. Besides this, the Cramer's V values of those five statistically values are also all above .25, which means that the associations are very strong.

Looking at the hypotheses, H1a is rejected because the indicators 'well-presented' and 'professional' are not significantly associated with each other in Table 19 since the p-value is above $\alpha = .05$. H1e is also rejected because there is no statistically significant association between 'no-spin' and 'concise' in both tables. On the other hand, H1b, H1c, H1d and H1f are all accepted since the tested indicators are significantly associated with each other since the p-values are below $\alpha = .05$. Furthermore, the indicators that identify rating bias and risk aversion are not significantly associated with each other due to the contrast in answers. On the other hand, the indicators that identify decoy effects and popularity bias are significantly associated with each other. The statistically significant association means that the indicators are mutually dependent. In the case of Amazon, most of the participants who answered that the indicator 'complete' does not contribute to the credibility of Amazon's recommender system, also answered that the indicators 'objective' and 'representative' do not contribute to the credibility. In the case of Tweakers, most of the participants who answered that the indicator 'complete' contributes to the credibility of Tweakers' recommender system, also answered that the indicators 'objective' and 'representative' contribute to the credibility. The 3x3 contingency tables with the observed and expected counts are presented in Appendix VI.

Amazon	LR χ^2	df	Significance χ^2 *$\alpha = .05$	Cramer's V	Significance Cramer's V *$\alpha = .05$
Well-presented * Professional	7.858	4	.097	.223	.085
Complete * Objective	15.198	4	.004*	.293	.007*
Complete * Representative	24.757	4	> .001*	.372	> .001*
Objective * Representative	18.686	4	.001*	.341	.001*
No-spin * Concise	7.861	4	.097	.217	.101
Complete * Expert	18.575	4	.001*	.339	.001*

Table 19. LR χ^2 values Amazon.

Tweakers	LR χ^2	df	Significance χ^2 *$\alpha = .05$	Cramer's V	Significance Cramer's V *$\alpha = .05$
Well-presented * Professional	18.883	4	.001*	.345	.001*
Complete * Objective	16.431	4	.002*	.320	.002*
Complete * Representative	27.490	4	> .001*	.410	> .001*
Objective * Representative	15.314	4	.004*	.323	.002*
No-spin * Concise	7.907	4	.095	.207	.134
Complete * Expert	9.534	4	.049*	.255	.030*

Table 20. LR χ^2 values Tweakers.

5.3 Post-test

The credibility of Amazon's recommender system in the post-test is lower ($M = 12.94$, $SD = 3.602$) than the credibility in the pre-test ($M = 13.91$, $SD = 3.422$). This decrease is statistically significant since the p-value is below $\alpha = .05$. On the contrary, the credibility of Tweakers' recommender system in the post-test is higher ($M = 14.07$, $SD = 3.254$) than the credibility in the pre-test ($M = 13.22$, $SD = 3.975$). This increase is also statistically significant since the p-value is below $\alpha = .05$. Thus, H2a and H2b are both accepted. The effect size $r = .31$ has a medium to large effect and $r = .24$ has a small to medium effect, according to the population effect size index of Cohen (1992). In the post-test, 45 participants provided a lower credibility score to Amazon's recommender system, while 41 participants provided a higher credibility score to Tweakers' recommender system. The results, with the scores per reflective indicator, are reported in Table 21 and Table 22. Table 23 reports the number of participants who have increased, decreased, or tied their credibility scores for both RS in the post-test.

Amazon	M post-test	SD	t	df	Significance *$\alpha = .05$	Effect size
Perceived accuracy	4.54	1.259	1.045	81	.115	$r = .12$
Perceived authenticity	4.11	1.474	4.981	81	> .001*	$r = .48$
Perceived believability	4.29	1.319	.992	81	.162	$r = .11$
Credibility	12.94	3.602	2.919	81	.003*	$r = .31$

Table 21. Post-test results Amazon.

Tweakers	M post-test	SD	t	df	Significance *$\alpha = .05$	Effect size
Perceived accuracy	4.84	1.153	-1.728	81	.044*	$r = .19$
Perceived authenticity	4.65	1.221	-1.491	81	.070	$r = .16$
Perceived believability	4.60	1.265	-2.585	81	.006*	$r = .28$
Credibility	14.07	3.254	-2.257	81	.014*	$r = .24$

Table 22. Post-test results Tweakers.

Credibility score changes	N	Increases	Decreases	Ties
Amazon	82	22	45	15
Tweakers	82	41	19	22

Table 23. Post-test changes.

Table 24 summarises the results of the hypothesis tests. H1 cannot be entirely accepted since H1a and H1d are rejected. H2 is entirely accepted.

Hypotheses	Accepted or rejected
H1a	Rejected
H1b	Accepted
H1c	Accepted
H1d	Accepted
H1e	Rejected
H1f	Accepted
H2a	Accepted
H2b	Accepted

Table 24. Summary hypothesis tests.

5.4 Reliability

Cronbach's α is applied to measure the reliability of the credibility scale. The results are reported in Table 25. The results of the Cronbach's α indicate that the credibility scale has acceptable scale reliability and internal consistency as Cronbach's α is $> .8$ in the pre-test and post-test. Moreover, the inter-item correlations were higher than the average of .57 for a three-item scale, which was indicated by Cortina (1993). The inter-item correlations of the three items of the measurements of this experiment vary from .62 to .88 (Appendix VII).

Credibility	Items	Cronbach's α
Pre-test credibility scale Amazon	3	.88
Pre-test credibility scale Tweakers	3	.93
Post-test credibility scale Amazon	3	.86
Post-test credibility scale Tweakers	3	.87

Table 25. Results of Cronbach's α .

5.5 Validity

The paper of Slack and Draugalis Jr. (2001) is used to assess the internal validity of this experiment. They state that there are eight threats to internal validity within research: History, maturation, testing, instrumentation, regression, selection, experimental validity, and interaction of threats. History and maturation are more of a concern in longitudinal studies. Testing becomes a threat in pre- and post-test designs when participants learn to provide the right answers after similar questions are replicated in a questionnaire. This could have been a threat within this study if participants firstly judged the credibility of Amazon's recommender system with the BOIT before judging Tweaker's recommender system in the pre-test. However, this threat has been avoided because participants judged both RS firstly without interacting and applying the BOIT. Instrumentation becomes a threat if results are due to changes in the instrument or measurement instead of a true treatment effect. This threat has been avoided since the measurements have not changed throughout the distribution period of the experiment. The regression threat can take place when subjects have been selected based on extreme scores. This is not the case within this experiment since the whole sample of $N = 82$ is used in the data analysis. The selection and experimental mortality threats can occur in between-subjects experimental designs. Thus, it is not a threat to this study. The last threat to internal validity is an interaction with the selection threat with other threats. Since selection does not threaten the internal validity of this study, the last threat is also avoided. Looking at external validity, the majority of the sample is younger than 26 and has followed a study on at least a bachelor level. Due to this, the findings correspond more to younger individuals that are highly educated.

Looking at content validity, it can be said that all the aspects of the concept of measuring credibility are covered, due to the usage of the most relevant triangulation methods, and the relevant indicators. They were all aligned with the relevant types of bias and overspecialisation. The credibility scale that was used for this experiment suggests high content quality because the three reflective indicators, accuracy, authenticity, and believability, make sense in the context of the definition of message credibility: an individual's judgment of the veracity of the content of the communication (Appelman & Sundar, 2016). Additionally, more participants answered that the formative indicators contribute to the credibility of Tweakers' recommender system than the one of Amazon. Additionally, the recommender system of Tweakers was also awarded a higher credibility score in the post-test. This resulted in a statistically significant transition of the most credible recommender system in the post-test. Hence, it can be argued that the results from the BOIT usage correspond to the credibility scores. This suggests high criterion validity.

Furthermore, the theory chapter revealed that the recommender system of Amazon lacks objectivity, expertise, representativeness, and fairness. It was also claimed that the recommender system of Tweakers lacks consistency. Looking at the results of the BOIT usage, most of the participants share those ideas since the results indicate that Tweakers' recommender system was more objective, expert, representative and fair. Additionally, fewer participants answered that the indicator 'consistent' contributed to the credibility of the recommender system of Tweakers compared to the recommender system of Amazon (Figure 5). Moreover, the reflective indicator 'authenticity' was defined as "recommendations that are bias-free and are therefore trusted" (Table 17). Looking at the usage of the triangulators and especially at the scores within the triangulators that are aligned with the types of bias, it was noticeable that the recommender system of Amazon received fewer 'yes-scores' and more 'no-scores' than that of Tweakers. This resulted in a statistically significant decrease in the score of the perceived authenticity of Amazon's recommender system. Therefore, it can be said that the operationalisation measures the concept that it is supposed to measure and thus, it suggests high construct validity.

6. Discussion & conclusion

Within this chapter, the key findings of the study will be reported, and the central research question will be answered. Next, the limitations, ideas for future research, and the implications of this study will be presented.

6.1 Key findings

This study aims to provide users with a mechanism that helps them with identifying rating bias, serial position effects, decoy effects, risk aversion, popularity bias and overspecialisation within RS in several application domains. Because of this, users can judge the credibility of RS more easily and decide if they want to neutralise the recommender system to avoid manipulation and irrelevant decisions. The central research question of this study was stated as: ***“What are the effects of the BOIT on users’ judgment of the credibility of recommender systems?”*** After the creation of the BOIT, conducting the experiment, and analysing the collected data, the following key findings can be reported.

The BOIT was created by applying the message credibility theory (Appelman & Sundar, 2016), the triangulation theory (Wijnhoven & Brinkhuis, 2015) in order to identify the types of bias and overspecialisation that were discussed within the problem analysis. The collaborative recommender system of the commercial organisation Amazon and the hybrid recommender system of the independent organisation Tweakers were used in an online within-subjects experiment with a pre-test and post-test. The experiment tested if certain formative indicators were significantly associated with each other and if the BOIT had a statistically significant effect on the users’ judgment of the credibility of RS in the post-test. In the pre-test, the recommender system of Amazon had a higher credibility score than that of Tweakers.

The results of the BOIT usage indicate that more participants have answered that the formative indicators do not contribute to the credibility of Amazon’s recommender system compared to that of Tweakers. The difference in answers between the indicators ‘representative’, ‘objective’, ‘expert’, and ‘fair’ are noticeable as a higher number of participants answered that these indicators do not contribute to the credibility of Amazon’s recommender system. Thus, it can be argued that the participants shared the ideas from the theory chapter. Within the theory chapter, it was claimed that the recommender system of Amazon lacks objectivity, expertise, representativeness, and fairness since it only recommends items from one brand with high ratings. Some of these items do not even fulfil the search requirements of the users in the scenario of the experiment. In addition, users of the recommender system of Amazon can be trapped in filter bubbles since they only receive recommendations from only ‘one point of view’: televisions from the brand Samsung. Due to this, there is evidence that the users can identify potential decoy effects, popularity bias, and overspecialisation within the recommender system of Amazon by applying the BOIT. Furthermore, the results from the BOIT usage reveal that more participants answered that the formative indicators contribute to the credibility of Tweakers’ recommender system. The recommender system of Tweakers recommends a more diverse range of items with different brands that fulfil the requirements of users instead of recommending items from only one brand that do not fulfil the requirements in the scenario.

In the post-test, 45 of the 82 participants provided the recommender system of Amazon compared with lower credibility scores. This decrease is statistically significant. Moreover, 41 of the 82 participants provided the recommender system of Tweakers with higher credibility scores. This increase is also statistically significant. For the one of Amazon, the reflective indicator with the main downgrade is ‘authenticity’. Most of the participants were able to identify decoy effects and popularity bias within the recommender system of Amazon. This led to a statistically significant decline in perceived authenticity as an authentic recommendation is defined as bias-free and trusted (Table 17). For the recommender system of Tweakers, the reflective indicator with the main statistically significant upgrade is ‘believability’. After identifying the types of bias and overspecialisation within the recommender system of Amazon, most of the participants did not find the same types of bias and overspecialisation within Tweakers’ recommender system. This could be a reason for the contribution to a higher score of perceived believability. However, this cannot be said with full certainty since the believability indicator is entirely subjective.

Moreover, the Cronbach's α values were all higher than $\alpha = .8$ in the pre-test and the post-test. Thus, the applied three-item credibility scale is considered reliable. The eight threats for internal validity, which are discussed by Slack and Draugalis Jr. (2001), were avoided since internal validity is crucial for studies that test the effectiveness of a mechanism and that aim to deliver replicable and robust results (Schram, 2005). Looking at measurement validity, it can be said that all the three reflective indicators, accuracy, authenticity, and believability, make sense in the context of the definition of message credibility. Besides, the formative indicators that were used to identify decoy effects and popularity bias and overspecialisation within the recommender system of Amazon are significantly associated with each other. An LR χ^2 test was conducted to test the level of association within those indicators. The Cramer's V values revealed that the strength of these associations is very strong. This validated the BOIT and indicated that the operationalisation measures the concept that it is supposed to measure.

The key findings of this study have now been presented. Therefore, the central research question can now be answered. The BOIT has affected the user's judgment of the credibility of RS in such a way that the recommender system of a commercial organisation became less credible than the one of an independent organisation in the post-test. Without applying the BOIT, all of the users were not able to identify the decoy effects, popularity bias, and overspecialisation within the recommender system of Amazon. Due to this, it can be concluded that the BOIT spreads awareness among users about potential biases and overspecialisation within RS and can therefore decrease the possibility of manipulation and irrelevant decisions.

6.2 Limitations and future research

Though this study is considered reliable and valid, several limitations still need to be addressed. This study's main focus was to test and validate the BOIT to see if it functions in a way that matches the statements made in the theory section. Mainly due to this, no specific research on extraneous variables within this study has been conducted. For this reason, it is not certain what participants' rating intentions were of the subjective 'believability' indicator. In addition, it was also not clear why some participants did not change their opinion or had a divergent opinion compared to the means. Because of the privacy reasons addressed in the methodology chapter, it was not possible to ask these participants again for the reasons for their answers. Subsequently, there is also no information regarding their real-life usage of RS and information regarding their brand preferences. Moreover, only two e-commerce RS were used within this study and it is not clear (yet) if the BOIT will have similar effects on the judgment of the credibility of other RS in different application domains or on other RS with different filtering algorithms.

Due to these limitations, the following ideas for future research are suggested. First, a similar research design with more specific questions regarding RS usage and brand preferences need to be conducted to test if these variables are extraneous. Besides, participants who have not changed their minds or who have divergent views in the post-test compared with the mean should receive additional questions regarding their choices. Second, it is suggested to use recommendation lists of RS from other application domains such as e-learning or e-tourism to test if the effects of the BOIT on users' judgment of the credibility of RS are similar to the results of this study.

6.3 Implications

The previous section suggests that the BOIT needs to be tested multiple times in different environments to prove that it still has a statistically significant effect on users' judgment of the credibility of RS. Though, it can be said that the first step regarding spreading awareness among users about bias and overspecialisation within RS has been made. This study discussed the types and bias and overspecialisation within RS and a BOIT was developed in a response to the key findings of Teppan and Zanker (2015). Teppan and Zanker (2015) argued in their paper that users need to be provided by a mechanism that allows them to identify and neutralise disingenuous biases in order to release persuasive power. Additionally, this study proposes new definitions to the three reflective indicators 'accuracy', 'authenticity' and 'believability' in the context of RS credibility. These reflective indicators are defined in such a way that they can determine the credibility of RS in various application domains to avoid manipulation and poor item decisions. Besides, this study proposes that the formative indicators and triangulators can be used as a set of requirements for relevant, serendipitous, diverse, bias-free, and trusted RS. Finally, the creation of an online application that consists of the BOIT can be realised if the suggested ideas for future research also reveal conclusions that are similar to this study. This online application needs to be freely available on computers and mobile devices. Therefore, users will be able to identify biases and overspecialisation in the same way as the participants did within this study in order to release the persuasive power of manipulated recommendations and to avoid irrelevant decisions in real life.

Reference list

- Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 42–46). <https://doi.org/10.1145/3109859.3109912>
- Abdollahpouri, H., Burke, R., & Mobasher, B. (2019). Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. *Eprint ArXiv:1901.07555*. Retrieved from <https://arxiv.org/abs/1901.07555>
- Adamopoulos, P., & Tuzhilin, A. (2015). On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4), 54. <https://doi.org/10.1145/2559952>
- Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546. <https://doi.org/10.1017/S0003055401003100>
- Adomavicius, G., Bockstedt, J., Curley, S. P., Zhang, J., & Ransbotham, S. (2019). The hidden side effects of recommendation systems. *MIT Sloan Management Review*, 60(2), 13–15. Retrieved from <https://sloanreview.mit.edu/article/the-hidden-side-effects-of-recommendation-systems/>
- Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2019). Reducing Recommender Systems Biases: An Investigation of Rating Display Designs. *Forthcoming, MIS Quarterly*, 19–18. <https://doi.org/10.25300/MISQ/2019/13949>
- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, 17(6), 734–749. <https://doi.org/10.1109/TKDE.2005.99>
- Aggarwal, C. C. (2016). An introduction to recommender systems. In *Recommender systems* (pp. 1–28). Springer. https://doi.org/10.1007/978-3-319-29659-3_1
- Ahmadian, S., Joorabloo, N., Jalili, M., Ren, Y., Meghdadi, M., & Afsharchi, M. (2020). A social recommender system based on reliable implicit relationships. *Knowledge-Based Systems*, 192, 105371. <https://doi.org/10.1016/j.knosys.2019.105371>
- Akoglu, H. (2018). User’s guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Amazon. (2020). Samsung Q85R 138 cm (55 inch) 4K QLED TV (Q HDR, Ultra HD, HDR, Twin Tuner, Smart TV). Retrieved February 11, 2020, from https://www.amazon.de/dp/B07Q62461Y/ref=sr_1_1?__mk_nl_NL=ÅMÅŽÕÑ&keywords=Samsung+55Q85R&qid=1580814960&s=ce-de&sr=1-1
- Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1), 59–79. <https://doi.org/10.1177/1077699015606057>
- Arazy, O., Kumar, N., & Shapira, B. (2010). A theory-driven design framework for social recommender systems. *Journal of the Association for Information Systems*, 11(9), 455. <https://doi.org/10.17705/1jais.00237>
- Badran, M., Bou abdo, J., Al Jurdi, W., & Demerjian, J. (2019). Adaptive Serendipity for Recommender Systems: Let It Find You. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 739–745. <https://doi.org/10.5220/0007409507390745>
- Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, 36(3), 421–458. <https://doi.org/10.2307/2393203>

- Bhasin, H. (2019). Top E-commerce companies in the world. Retrieved January 7, 2020, from <https://www.marketing91.com/e-commerce-companies-in-the-world/>
- BMS. (2020). Qualtrics. Retrieved January 23, 2020, from <https://www.utwente.nl/en/bms/datalab/datacollection/surveysoftware/qualtrics/>
- Bobadilla, J., Ortega, F., Hernando, A., & Alcalá, J. (2011). Improving collaborative filtering recommender system results and performance using genetic algorithms. *Knowledge-Based Systems*, 24(8), 1310–1316. <https://doi.org/10.1016/j.knosys.2011.06.005>
- Boone, H. N., & Boone, D. A. (2012). Analyzing Likert Data. *Journal of Extension*, 50(2), 1–5. Retrieved from <https://www.joe.org/joe/2012april/tt2.php>
- Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4), 249–265. <https://doi.org/10.1007/s10676-015-9380-y>
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 1–38. <https://doi.org/10.5334/joc.72>
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370. <https://doi.org/10.1023/A:1021240730564>
- Burke, R., Felfernig, A., & Göker, M. H. (2011). Recommender systems: An overview. *Ai Magazine*, 32(3), 13–18. <https://doi.org/10.1609/aimag.v32i3.2361>
- Burke, R., & Ramezani, M. (2011). Matching recommendation technologies and domains. In *Recommender systems handbook* (pp. 367–386). Springer. https://doi.org/1007/978-0-387-85820-3_11
- Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6), 1487–1524. <https://doi.org/10.3233/IDA-163209>
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8. <https://doi.org/10.1016/j.jebo.2011.08.009>
- Churchman, C. W. (1971). *The Design of Inquiring Systems Basic Concepts of Systems and Organization*. Basic Books.
- Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112(1), 155. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98. <https://doi.org/10.1037/0021-9010.78.1.98>
- Courtney, J. F. (2001). Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for DSS. *Decision Support Systems*, 31(1), 17–38. [https://doi.org/10.1016/S0167-9236\(00\)00117-2](https://doi.org/10.1016/S0167-9236(00)00117-2)
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- de Gemmis, M., Lops, P., Semeraro, G., & Musto, C. (2015). An investigation on the serendipity problem in recommender systems. *Information Processing & Management*, 51(5), 695–717. <https://doi.org/10.1016/j.ipm.2015.06.008>
- Denzin, N. K. (2015, October 26). Triangulation. *The Blackwell Encyclopedia of Sociology*. <https://doi.org/10.1002/9781405165518.wbeost050.pub2>
- DePillis, L., & Sherman, I. (2019). Amazon’s extraordinary 25-year evolution. Retrieved January 7,

2020, from <https://edition.cnn.com/interactive/2018/10/business/amazon-history-timeline/index.html%0D>

- Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative Filtering Recommender Systems. *Foundations and Trends® in Human-Computer Interaction*, 4(2), 81–173. <https://doi.org/10.1561/1100000009>
- Fatemi, M., & Tokarchuk, L. (2013). A community based social recommender system for individuals & groups. In *2013 International Conference on Social Computing* (pp. 351–356). IEEE. <https://doi.org/10.1109/SocialCom.2013.55>
- Felfernig, A., Friedrich, G., Gula, B., Hitz, M., Kruggel, T., Leitner, G., ... Teppan, E. (2007). Persuasive recommendation: serial position effects in knowledge-based recommender systems. In *International Conference on Persuasive Technology* (pp. 283–294). Springer. https://doi.org/10.1007/978-3-540-77006-0_34
- Field, A. (2009). *Discovering statistics using SPSS*. SAGE Publications Limited.
- Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1), 298–320. <https://doi.org/10.1093/poq/nfw006>
- Fusch, P., Fusch, G. E., & Ness, L. R. (2018). Denzin's paradigm shift: Revisiting triangulation in qualitative research. *Journal of Social Change*, 10(1), 2. <https://doi.org/10.5590/JOSC.2018.10.1.02>
- Google. (2020). 55-inch televisie. Retrieved April 3, 2020, from shorturl.at/bfIW8
- Guo, G., Zhang, J., & Thalmann, D. (2014). Merging trust in collaborative filtering to alleviate data sparsity and cold start. *Knowledge-Based Systems*, 57, 57–68. <https://doi.org/10.1016/j.knosys.2013.12.007>
- Huang, S. (2011). Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods. *Electronic Commerce Research and Applications*, 10(4), 398–407. <https://doi.org/10.1016/j.elerap.2010.11.003>
- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Enhancement of the Neutrality in Recommendation. In *Decisions@ RecSys* (pp. 8–14). Retrieved from https://www.researchgate.net/publication/285911417_Enhancement_of_the_Neutrality_in_Recommendation
- Kaptein, M., Markopoulos, P., De Ruyter, B., & Aarts, E. (2015). Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77, 38–51. <https://doi.org/10.1016/j.ijhcs.2015.01.004>
- Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Technical Report EBSE 2007-001*. <https://doi.org/10.1.1.117.471>
- Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111, 180–192. <https://doi.org/10.1016/j.knosys.2016.08.014>
- Kouki, P., Fakhraei, S., Foulds, J., Eirinaki, M., & Getoor, L. (2015). Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 99–106). ACM. <https://doi.org/10.1145/2792838.2800175>
- Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems*, 74, 12–32. <https://doi.org/10.1016/j.dss.2015.03.008>
- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1), 151–169.

<https://doi.org/10.1146/annurev.publhealth.23.100901.140546>

- Madadipouya, K., & Chelliah, S. (2017). A Literature Review on Recommender Systems Algorithms, Techniques and Evaluations. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 8(2), 109–124. Retrieved from <http://www.brain.edusoft.ro/index.php/brain/article/view/693>
- Maksai, A., Garcin, F., & Faltings, B. (2015). Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 179–186). <https://doi.org/10.1145/2792838.2800184>
- Mason, R. O., & Mitroff, I. I. (1973). A program for research on management information systems. *Management Science*, 19(5), 475–487. <https://doi.org/10.1287/mnsc.19.5.475>
- Matt, C., Benlian, A., Hess, T., & Weiß, C. (2014). Escaping from the filter bubble? The effects of novelty and serendipity on users' evaluations of online recommendations. *Proceedings of the 35th International Conference on Information Systems*, 1–18.
- Mchugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. <https://doi.org/10.11613/bm.2013.018>
- Milano, S., Taddeo, M., & Floridi, L. (2019). Recommender Systems and their Ethical Challenges. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3378581>
- Montaner, M., López, B., & De La Rosa, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, 19(4), 285–330. <https://doi.org/10.1023/A:1022850703159>
- Nagulendra, S., & Vassileva, J. (2014). Understanding and controlling the filter bubble through interactive visualization: a user study. In *Proceedings of the 25th ACM conference on Hypertext and social media* (pp. 107–115). ACM. <https://doi.org/10.1145/2631775.2631811>
- O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces* (pp. 167–174). ACM. <https://doi.org/10.1145/1040830.1040870>
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- Park, D.-H., Lee, J., & Han, I. (2006). Information overload and its consequences in the context of online consumer reviews. *PACIS 2006 Proceedings*, 28. Retrieved from https://www.researchgate.net/publication/221228936_Information_Overload_and_its_Consequences_in_the_Context_of_Online_Consumer_Reviews
- Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5–6), 393–408. <https://doi.org/10.1023/A:1006544522159>
- Ricci, F., Kantor, P. B., Rokach, L., & Shapira, B. (2011). *Recommender Systems Handbook*. Springer. https://doi.org/10.1007/978-0-387-85820-3_1
- Schafer, J. Ben, Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291–324). Springer. https://doi.org/10.1007/978-3-540-72079-9_9
- Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2), 225–237. <https://doi.org/10.1080/13501780500086081>
- Simran, S., Pande, A., & Desai, P. (2019). Preference-Search based Recommendation System for Accommodation Facilitator: A Review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 5(2), 951–956. <https://doi.org/10.32628/CSEIT1952245>

- Singh, M., & Mehrotra, M. (2016). Bridging the Gap Between Users and Recommender Systems: A Change in Perspective to User Profiling. In *Intelligent Systems Technologies and Applications* (pp. 379–386). Springer. https://doi.org/10.1007/978-3-319-23258-4_33
- Slack, M. K., & Draugalis Jr, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy*, 58(22), 2173–2181. <https://doi.org/10.1093/ajhp/58.22.2173>
- Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon. com. *Ieee Internet Computing*, 21(3), 12–18. <https://doi.org/10.1109/MIC.2017.72>
- Steckler, A., & Mcleroy, K. R. (2008). The Importance of External Validity. *American Journal of Public Health*, 98(1), 9–10. <https://doi.org/10.2105/ajph.2007.126847>
- Sudesh, G., Dharmic, B., Pulari, S. R., & Ramesh, G. (2018). An extensive study on various methodologies used in the field of recommender systems. *International Journal of Pure and Applied Mathematics*, 119(18), 1961–1970.
- Taken, F. (2008). De geschiedenis van Tweakers.net: ontstaan en hobbyfase. Retrieved January 8, 2020, from <https://tweakers.net/reviews/991/de-geschiedenis-van-tweakers-punt-net-ontstaan-en-hobbyfase.html>
- Teppan, E. C., & Felfernig, A. (2012). Minimization of decoy effects in recommender result sets. *Web Intelligence and Agent Systems: An International Journal*, 10(4), 385–395. <https://doi.org/10.3233/WIA-2012-0253>
- Teppan, E. C., & Zanker, M. (2015). Decision Biases in Recommender Systems. *Journal of Internet Commerce*, 14(2), 255–275. <https://doi.org/10.1080/15332861.2015.1018703>
- Tweakers. (2020a). Over Tweakers, Onafhankelijkheid. Retrieved January 8, 2020, from <https://tweakers.net/info/over-tweakers/onafhankelijkheid/>
- Tweakers. (2020b). Samsung 55Q85R Zwart. Retrieved February 11, 2020, from <https://tweakers.net/pricewatch/1329778/samsung-55q85r-zwart/alternatieven/>
- Victor, P., Cornelis, C., & De Cock, M. (2011). *Trust networks for recommender systems* (Vol. 4). Springer Science & Business Media.
- Wijnhoven, F., & Brinkhuis, M. (2015). Internet information triangulation: Design theory and prototype evaluation. *Journal of the Association for Information Science and Technology*, 66(4), 684–701. <https://doi.org/10.1002/asi.23203>
- Wood, P. K. (1983). Inquiring systems and problem structure: Implications for cognitive development. *Human Development*, 26(5), 249–265. <https://doi.org/10.1159/000272887>
- Zhou, X., Xu, Z., Sun, X., & Wang, Q. (2017). A new information theory-based serendipitous algorithm design. In *International Conference on Human Interface and the Management of Information* (pp. 314–327). Springer. https://doi.org/10.1007/978-3-319-58524-6_26
- Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should We Worry about Filter Bubbles? *Internet Policy Review*, 5(1). <https://doi.org/10.14763/2016.1.401>

Appendices

Appendix I: Paper selection procedure for chapter 2: Problem analysis

Source	URL
Google Scholar	https://scholar.google.nl/
Scopus	https://www.scopus.com/

Inclusion criteria	Exclusion criteria
Papers presenting the searched keywords.	Papers not presenting the searched keywords at all.
Papers from conferences and journals.	Papers that present RS are purely and only written from the Computer Science perspective addressing codes, formula's and language that are not relevant to a Business Administration perspective.
Papers are written in English.	Papers that do not provide detailed information.
Papers that can be used in a Business Administration perspective.	Grey literature.
Scientific books.	Papers that provide information about unfinished research.










Keywords
1. Recommender systems
2. Recommender systems bias
3. Recommender systems overspecialisation

Keywords	Search and retrieval	Coarse selection	Detailed selection (Used papers in chapter 2)
1.	197,115	35	22
2.	35,845	12	9
3.	4,230	23	13
Total	237,190	70	44

Appendix II: Sponsored recommendations

55-inch tel... bekijken Gesponsord ⓘ

[Producten](#) [Vergelijkingssites](#)

 Philips 43PFS5503/12 € 279,00 Kamera Express ★★★★★ (65) Van beslist.nl	 PHILIPS Philips 43PFS5503 televisie € 329,00 Coolblue Gratis verzending Van beslist.nl	 Philips 50PUS6704/12 € 499,00 MediaMarkt ★★★★★ (572) Van beslist.nl
 SAMSUNG Samsung UE32N5000 televisie € 299,00 Coolblue Gratis verzending Van beslist.nl	 SAMSUNG Samsung UE55RU7170 € 479,00 Kamera Express ★★★★★ (158) Van beslist.nl	 SALE Toshiba 55U2963DG led-tv (139 cm / 55 inc...) € 399,99 € 699 OTTO Van Shoptail
 SAMSUNG Samsung UE50RU7020 televisie € 449,00 Coolblue Gratis verzending Van beslist.nl	 Philips 24PFS5863 nieuwstaat... € 251,00 Gereviseerd PlatteTV Van Bigshopper	 SALORA Salora 55UHL2800 € 382,00 Kijkshop.nl Gratis verzending Van Google

(Google, 2020)

These recommendations are sponsored, which can be seen in the top right corner. This results in poor recommendations because the first four items do not even have the 55-inch requirement. In addition, the first three items are from the same brand. The ratings are only displayed for three of the nine items.

Appendix III: Rating types within RS



(Adomavicius et al., 2019, p.12)

Appendix IV: Demographic data of the sample

Gender	Frequency	Per cent
Male	39	47.6
Female	43	52.4
Total	82	100.0

Age	
N	82
Mean	25.6707
Standard deviation	9.72086
Minimum	18
Maximum	55

Age	Frequency	Per cent	Cumulative per cent
18	4	4.9	4.9
19	11	13.4	18.3
20	10	12.2	30,5
21	10	12.2	42.7
22	10	12.2	54.9
23	7	8.5	63.4
24	5	6.1	69.5
25	5	6.1	75,6
26	3	3.7	79.3
27	2	2.4	81.7
28	1	1.2	82.9
29	1	1.2	84.1
31	1	1.2	85,4
33	1	1.2	86.6
38	1	1.2	87.8
42	1	1.2	89.0
46	1	1.2	90.2
47	2	2.4	92.7
49	1	1.2	93.9
51	1	1.2	95.1
52	1	1.2	96.3

54	2	2.4	98.8
55	1	1.2	100.0
Total	82	100.0	

Highest followed education	Frequency	Per cent
Less than high school graduate	1	1.2
Associate degree	12	13.8
Bachelor's degree	63	76.8
Master's degree	6	7.3
Doctorate	0	0
Total	82	100.0

Nationality	Frequency	Per cent
Dutch	38	46.3
German	36	43.9
Turkish	4	4.9
Italian	2	2.4
Latvian	1	1.2
Lithuanian	1	1.2
Total	82	100

Appendix V: Usage of BOIT

Data triangulator, Amazon	Yes	Maybe	No	Total
Complete	32 (39%)	27 (32.9%)	23 (28%)	82 (100%)
Representative	29 (35.4%)	25 (30.5%)	28 (34.1%)	82 (100%)
Total scores	61 (37.2%)	52 (31.7%)	51 (31.1%)	164 (100%)

Theory triangulator, Amazon	Yes	Maybe	No	Total
Consistent	43 (52.4%)	26 (31.7%)	13 (15.9%)	82 (100%)
Concise	32 (39%)	34 (41.5%)	16 (19.5%)	82 (100%)
Well-presented	45 (54.9%)	19 (23.2%)	18 (22%)	82 (100%)
Total	120 (48.8%)	79 (32.1%)	47 (19.1%)	246 (100%)

Investigator triangulator, Amazon	Yes	Maybe	No	Total
Objective	17 (20.7%)	24 (29.3%)	41 (50%)	82 (100%)
No-spin	14 (17.1%)	48 (58.5%)	20 (24.4%)	82 (100%)
Expert	21 (25.6%)	23 (28%)	38 (46.3%)	82 (100%)
Total	52 (21.1%)	95 (38.6%)	99 (40.3%)	246 (100%)

Methods triangulator, Amazon	Yes	Maybe	No	Total
Professional	34 (41.5%)	21 (25.6%)	27 (32.9%)	82 (100%)
Total	34 (41.5%)	21 (25.6%)	27 (32.9%)	82 (100%)

Relevance triangulator, Amazon	Yes	Maybe	No	Total
Will have impact	29 (35.4%)	39 (47.6%)	14 (17.1%)	82 (100%)
Fair	20 (24.4%)	32 (39%)	30 (36.6%)	82 (100%)
Ease of use	50 (61%)	23 (28%)	9 (11%)	82 (100%)
Total	99 (40.3%)	94 (38.2%)	53 (21.5%)	246 (100%)

Data triangulator, Tweakers	Yes	Maybe	No	Total
Complete	45 (54.9%)	22 (26.8%)	15 (18.3%)	82 (100%)
Representative	51 (62.2%)	21 (25.6%)	10 (12.2%)	82 (100%)
Total	96 (58.6%)	43 (26.2%)	25 (15.2%)	164 (100%)

Theory triangulator, Tweakers	Yes	Maybe	No	Total
Consistent	39 (47.6%)	36 (43.9%)	7 (8.5%)	82 (100%)
Concise	36 (43.9%)	33 (40.2%)	13 (15.9%)	82 (100%)
Well-presented	38 (46.3%)	27 (32.9%)	17 (20.7%)	82 (100%)
Total	113 (46%)	96 (39%)	37 (15.0%)	246 (100%)

Investigator triangulator, Tweakers	Yes	Maybe	No	Total
Objective	37 (45.1%)	24 (29.3%)	21 (25.6%)	82 (100%)
No-spin	23 (28%)	48 (58.5%)	11 (13.4%)	82 (100%)
Expert	34 (41.5%)	32 (39%)	16 (19.5%)	82 (100%)
Total	94 (38.2%)	104 (42.3%)	48 (19.5%)	246 (100%)

Methods triangulator, Tweakers	Yes	Maybe	No	Total
Professional	29 (35.4%)	31 (37.8%)	22 (26.8%)	82 (100%)
Total	29 (35.4%)	31 (37.8%)	22 (26.8%)	82 (100%)

Relevance triangulator, Tweakers	Yes	Maybe	No	Total
Will have impact	39 (47.6%)	37 (45.1%)	6 (7.3%)	82 (100%)
Fair	45 (54.9%)	31 (37.8%)	6 (7.3%)	82 (100%)
Ease of use	49 (59.8%)	28 (34.1%)	5 (6.1%)	82 (100%)
Total	133 (54.1%)	96 (39%)	17 (6.9%)	246 (100%)

Appendix VI: LR χ^2 3x3 contingency tables

Amazon Well-presented * Professional			Professional			Total
			1 Yes	2 Maybe	3 No	
Well-presented	1 Yes	Count	24	9	12	45
		Expected Count	18.7	11.5	14.8	45.0
	2 Maybe	Count	5	8	6	19
		Expected Count	7.9	4.9	6.3	19.0
	3 No	Count	5	4	9	18
		Expected Count	7.5	4.6	5.9	18.0
Total		Count	34	21	27	82
		Expected Count	34.0	21.0	27.0	82.0

Amazon Complete * Objective			Objective			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	8	12	12	32
		Expected Count	6.6	9.4	16.0	32.0
	2 Maybe	Count	8	9	10	27
		Expected Count	5.6	7.9	13.5	27.0
	3 No	Count	1	3	19	23
		Expected Count	4.8	6.7	11.5	23.0
Total		Count	17	24	41	82
		Expected Count	17.0	24.0	41.0	82.0

Amazon Complete * Representative			Representative			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	19	10	3	32
		Expected Count	11.3	9.8	10.9	32.0
	2 Maybe	Count	8	9	10	27
		Expected Count	9.5	8.2	9.2	27.0
	3 No	Count	2	6	15	23
		Expected Count	8.1	7.0	7.9	23.0
Total		Count	29	25	28	82
		Expected Count	29.0	25.0	28.0	82.0

Amazon Objective * Representative			Representative			Total
			1 Yes	2 Maybe	3 No	
Objective	1 Yes	Count	11	4	2	17
		Expected Count	6.0	5.2	5.8	17.0
	2 Maybe	Count	8	12	4	24
		Expected Count	8.5	7.3	8.2	24.0
	3 No	Count	10	9	22	41
		Expected Count	14.5	12.5	14.0	41.0
Total		Count	29	25	28	82
		Expected Count	29.0	25.0	28.0	82.0

Amazon No-spin * Concise			Concise			Total
			1 Yes	2 Maybe	3 No	
No-spin	1 Yes	Count	5	4	5	14
		Expected Count	5.5	5.8	2.7	14.0
	2 Maybe	Count	23	19	6	48
		Expected Count	18.7	19.9	9.4	48.0
	3 No	Count	4	11	5	20
		Expected Count	7.8	8.3	3.9	20.0
Total		Count	32	34	16	82
		Expected Count	32.0	34.0	16.0	82.0

Amazon Complete * Expert			Expert			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	13	7	12	32
		Expected Count	8.2	9.0	14.8	32.0
	2 Maybe	Count	6	13	8	27
		Expected Count	6.9	7.6	12.5	27.0
	3 No	Count	2	3	18	23
		Expected Count	5.9	6.5	10.7	23.0
Total		Count	21	23	38	82
		Expected Count	21.0	23.0	38.0	82.0

Tweakers Well-presented * Professional			Professional			Total
			1 Yes	2 Maybe	3 No	
Well-presented	1 Yes	Count	21	12	5	38
		Expected Count	13.4	14.4	10.2	38.0
	2 Maybe	Count	6	14	7	27
		Expected Count	9.5	10.2	7.2	27.0
	3 No	Count	2	5	10	17
		Expected Count	6.0	6.4	4.6	17.0
Total		Count	29	31	22	82
		Expected Count	29.0	31.0	22.0	82.0

Tweakers Complete * Objective			Objective			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	28	11	6	45
		Expected Count	20.3	13.2	11.5	45.0
	2 Maybe	Count	5	10	7	22
		Expected Count	9.9	6.4	5.6	22.0
	3 No	Count	4	3	8	15
		Expected Count	6.8	4.4	3.8	15.0
Total		Count	37	24	21	82
		Expected Count	37.0	24.0	21.0	82.0

Tweakers Complete * Representative			Representative			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	36	9	0	45
		Expected Count	28.0	11.5	5.5	45.0
	2 Maybe	Count	11	8	3	22
		Expected Count	13.7	5.6	2.7	22.0
	3 No	Count	4	4	7	15
		Expected Count	9.3	3.8	1.8	15.0
Total		Count	51	21	10	82
		Expected Count	51.0	21.0	10.0	82.0

Tweakers Objective * Representative			Representative			Total
			1 Yes	2 Maybe	3 No	
Objective	1 Yes	Count	29	6	2	37
		Expected Count	23.0	9.5	4.5	37.0
	2 Maybe	Count	13	10	1	24
		Expected Count	14.9	6.1	2.9	24.0
	3 No	Count	9	5	7	21
		Expected Count	13.1	5.4	2.6	21.0
Total		Count	51	21	10	82
		Expected Count	51.0	21.0	10.0	82.0

Tweakers No-spin * Concise			Concise			Total
			1 Yes	2 Maybe	3 No	
No-spin	1 Yes	Count	14	8	1	23
		Expected Count	10.1	9.3	3.6	23.0
	2 Maybe	Count	20	19	9	48
		Expected Count	21.1	19.3	7.6	48.0
	3 No	Count	2	6	3	11
		Expected Count	4.8	4.4	1.7	11.0
Total		Count	36	33	13	82
		Expected Count	36.0	33.0	13.0	82.0

Tweakers Complete * Expert			Expert			Total
			1 Yes	2 Maybe	3 No	
Complete	1 Yes	Count	23	16	6	45
		Expected Count	18.7	17.6	8.8	45.0
	2 Maybe	Count	8	11	3	22
		Expected Count	9.1	8.6	4.3	22.0
	3 No	Count	3	5	7	15
		Expected Count	6.2	5.9	2.9	15.0
Total		Count	34	32	16	82
		Expected Count	34.0	32.0	16.0	82.0

Appendix VII: Inter-item correlations credibility scale

Pre-test Amazon	Accuracy	Authenticity	Believability
Accuracy	1	.68	.73
Authenticity	.68	1	.72
Believability	.73	.72	1

Pre-test Tweakers	Accuracy	Authenticity	Believability
Accuracy	1	.78	.79
Authenticity	.78	1	.88
Believability	.79	.88	1

Post-test Amazon	Accuracy	Authenticity	Believability
Accuracy	1	.62	.62
Authenticity	.62	1	.80
Believability	.62	.80	1

Post-test Tweakers	Accuracy	Authenticity	Believability
Accuracy	1	.65	.65
Authenticity	.65	1	.79
Believability	.65	.79	1

Appendix VIII: Questionnaire

The Credibility of Recommender Systems

Dear participant,

First, thank you for taking the time to help me with this online experiment for my research. My name is Akansel Özgören and I am a master's student Business Administration at the University of Twente. This experiment will be conducted for my master thesis. I am conducting research on the credibility of recommender systems (RS).

This experiment will take approximately 10 to 15 minutes and will be entirely anonymous. You also have the possibility to withdraw from the experiment at any time you wish.

If you have any questions regarding the experiment or other cases, you can send me an email or call me if it is necessary.

Thank you and I wish you good luck with the experiment!

Kind regards,

Akansel Özgören
Master Student Business Administration
Phone: +31 570 671113
Mobile: +31 6 14582715
Email (private): akanselo zgoren@hotmail.com
Email (study): a.oezgoeren@student.utwente.nl

By ticking the box, you are accepting to be a participant of my research and you consent to the use of your answers for this research. This research is anonymous and will only be used as data for my master thesis.

☐ Yes, I consent.

☐ No, I do not consent.

Thank you for being a participant for my experiment! There are still some types of bias in multiple recommender systems (RS), also in the e-commerce world. This can decrease the credibility of RS and it can also persuade consumers in an unethical way. This can lead to wrong purchases and a waste of money. Nowadays, designers have not found any solution that can entirely fix this problem. However, consumers can recognise biases by themselves. But how?

The definition of credibility in this case is: "An individual's judgment of the veracity of the content of communication" and it can be measured with three subjective indicators: Accuracy (are the recommendations similar to the main product and relevant for you?), authenticity (do you trust the recommendations?) and believability (do you believe that the recommendations

are the ones you need?). On the next page, you will be exposed to a scenario.

In this scenario, you are looking for a new 55-inch television. You have found the same Samsung Q85R television on two different websites: Amazon and Tweakers, but you want to check the recommended products on both websites to look for alternatives. However, both websites show different recommendations. I will ask you to judge both RS on credibility, by only using your own opinion and the three subjective indicators.



(Amazon, 2020)

How credible do you think the recommender system of Amazon is? (Scale: 1: very poorly to 7: very well)

	1	2	3	4	5	6	7
Accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Authenticity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Believability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

TV en home cinema > Televisies

Gebruikers
★★★★★
6 reviews

Samsung 55Q85R Zwart

Prijz € 1.278,- ☐ prijs volgen

Specificaties 55" • Ultra HD • LCD • HDR • 2 tuners • Bekijk alle specificaties

Product uitvoering 55" Zwart

☐ willen ☐ hebben ☐ vergelijken

☐ Vergelijk

LG OLED55B8PLA Zwart

★★★★★ € 1.171,-

[Bekijk >](#)

☐ Vergelijk

LG OLED55B9PLA Zwart

★★★★★ € 1.008,-

[Bekijk >](#)

☐ Vergelijk

LG OLED55C8PLA Zwart

★★★★★ € 1.145,-

[Bekijk >](#)

☐ Vergelijk

LG OLED55C9PLA Zwart

★★★★★ € 1.398,94

[Bekijk >](#)

☐ Vergelijk

Philips 55OLED754 Zwart

€ 1.299,-

[Bekijk >](#)

☐ Vergelijk

Philips 55OLED803 Zwart

★★★★★ € 1.239,-

[Bekijk >](#)

☐ Vergelijk

Philips 55OLED854 Zwart

★★★★★ € 1.399,-

[Bekijk >](#)

☐ Vergelijk

Philips 55POS0002/12 Zwart

★★★★★ € 1.195,-

[Bekijk >](#)

☐ Vergelijk

Samsung QLED QE55Q8D Zwart

★★★★★ € 1.051,-

[Bekijk >](#)

☐ Vergelijk

Samsung QLED QE55Q80RAL Zilver, Zwart

★★★★★ € 1.129,-

[Bekijk >](#)

☐ Vergelijk

Sony Bravia KD-55XG8505 Zwart

★★★★★ € 1.195,-

[Bekijk >](#)

☐ Vergelijk

Sony KD-55AF8 Zwart

★★★★★ € 1.395,-

[Bekijk >](#)

(Tweakers, 2020b)

How credible do you think the recommender system of Tweakers is? (Scale: 1: very poorly to 7: very well)

	1	2	3	4	5	6	7
Accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Authenticity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Believability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

END OF PRE-TEST

POST-TEST

I have developed a checklist that can be used as a tool to make the judgment of the credibility of a recommender system easier. I will first provide you with information about the types of bias that are still present within RS. After this, I will ask you to have a look again at the two RS and rate them again, but then with applying my checklist.

Bias and overspecialisation	Definition
Rating bias	Poorly presented with numerical ratings and unprofessional (not based on studies or research).
Serial position effects	Inconsistent list with predetermined products in the beginning and the end of the list.
Decoy effects	Incomplete list with only predetermined products without competing products. This is not objective and not representative
Risk aversion	Less risky products with low utility. This is a type of spin manipulation and makes RS less concise .
Popularity bias	Incomplete list with only popular and well-known products. Expert RS include also less-known products from different brands.
Unexpectedness	Unexpected and unknown products that are useful will have an impact on the behaviour of consumers.
Filter bubble and echo chamber	When consumers receive an unfair selection of products from only one or two brands with strengthened persuasion.



Samsung Q85R 138 cm (55 inch) 4K QLED TV (Q HDR, Ultra HD, HDR, Twin Tuner, Smart TV)

Samsung
★★★★★ 8 beoordelingen | 6 beantwoorde vragen

Prijs: € 1.450,00
 Prijzen voor items die verkocht worden door Amazon zijn inclusief Duitse btw. Zie details voor andere items.

Nieuw (4) van € 1.450,00 + GRATIS verzending

- Samsung TV QLED 55" 138cm
- Ultra HD 4K





Samsung
GQ55Q85RGTXZG Flat
QLED TV Q85R (2019) 55
inch
★★★★★ 39
€ 1.499,00



Samsung Q80R 138 cm
(55 inch) 4K QLED TV
QE55Q80R (Q HDR, Ultra
HD, HDR, Twin Tuner,...
★★★★★ 2
€ 1.250,00



Samsung
GQ55Q80RGTXZG Flat
QLED TV Q80R (2019) 55
inch
★★★★★ 26
9 aanbiedingen van EUR
1.230,00



Samsung
GQ65Q85RGTXZG zwart
★★★★★ 39
€ 2.038,95



Samsung Q85R 163 cm
(65 inch) 4K QLED TV
65Q85R (Q HDR, Ultra HD,
HDR, Twin Tuner, Smart
TV) [energieklasse A+]
€ 1.880,00



Samsung Q90R 138 cm
(55 inch) 4K QLED TV (Q
HDR, Ultra HD, HDR, Twin
Tuner, Smart TV)
★★★★★ 3
€ 1.949,00



Samsung
GQ65Q90RGTXZG Qled-tv
(2019) 55 inch zwart
★★★★★ 32
10 aanbiedingen van EUR
1.399,00



Samsung Q85R 189 cm
(75 inch) 4K QLED TV (Q
HDR, Ultra HD, HDR, Twin
Tuner, Smart TV) ...
2 aanbiedingen van EUR
3.070,00



Samsung
GQ55Q70RGTXZG zwart
★★★★★ 68
6 aanbiedingen van EUR
958,38



Samsung Q80R 163 cm
(65 inch) 4K QLED TV
QE65Q80R (Q HDR, Ultra
HD, HDR, Twin Tuner,...
★★★★★ 1
€ 1.599,00



Samsung
GQ55Q80RGTXZG Flat
QLED TV Q80R (2019) 65
inch
★★★★★ 26
7 aanbiedingen van EUR
1.550,00



Samsung
GQ65Q90RGTXZG Qled-tv
(2019) 65 inch zwart
★★★★★ 32
€ 2.489,00

(Amazon, 2020)

Do these indicators contribute to the credibility of the recommender system of Amazon? (In the information table, the indicators are aligned with the types of bias and overspecialisation. The indicators are presented in bold italics. With 'ease of use', you need to check if the recommender system was easy to use or not.

	Yes	Maybe	No
Complete	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Representative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Consistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Concise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Well-presented	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Objective	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No spin manipulation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Professional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Will have impact	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ease of use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

After checking the indicators, please rate the credibility of the recommender system of Amazon again (Scale: 1: very poorly to 7: very well)

Reminder: Accuracy: Are the recommendations similar to the main product and relevant for you? Authenticity: Do you trust the recommendations? Believability: Do you believe that the recommendations are the ones you need?

	1	2	3	4	5	6	7
Accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Authenticity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Believability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Bias and overspecialisation	Definition
Rating bias	Poorly presented with numerical ratings and unprofessional (not based on studies or research).
Serial position effects	Inconsistent list with predetermined products in the beginning and the end of the list.
Decoy effects	Incomplete list with only predetermined products without competing products. This is not objective and not representative
Risk aversion	Less risky products with low utility. This is a type of spin manipulation and makes RS less concise .
Popularity bias	Incomplete list with only popular and well-known products. Expert RS include also less-known products from different brands.
Unexpectedness	Unexpected and unknown products that are useful will have an impact on the behaviour of consumers.
Filter bubble and echo chamber	When consumers receive an unfair selection of products from only one or two brands with strengthened persuasion.

TV en home cinema > Televisies

Gebruikers
★★★★★
6 reviews

Samsung 55Q85R Zwart

Prijz € 1.278,- ☐ prijs volgen

Specificaties 55" • Ultra HD • LCD • HDR • 2 tuners • [Bekijk alle specificaties](#)

Product uitvoering 55" Zwart

☐ willen ☐ hebben ☐ vergelijken

<input type="checkbox"/> Vergelijk LG OLED55B8PLA Zwart ★★★★★ € 1.171,- Bekijk »	<input type="checkbox"/> Vergelijk LG OLED55B9PLA Zwart ★★★★★ € 1.098,- Bekijk »	<input type="checkbox"/> Vergelijk LG OLED55C8PLA Zwart ★★★★★ € 1.145,- Bekijk »	<input type="checkbox"/> Vergelijk LG OLED55C9PLA Zwart ★★★★★ € 1.398,94 Bekijk »
<input type="checkbox"/> Vergelijk Philips 55OLED754 Zwart € 1.299,- Bekijk »	<input type="checkbox"/> Vergelijk Philips 55OLED803 Zwart ★★★★★ € 1.239,- Bekijk »	<input type="checkbox"/> Vergelijk Philips 55OLED854 Zwart ★★★★★ € 1.399,- Bekijk »	<input type="checkbox"/> Vergelijk Philips 55POS9002/12 Zwart ★★★★★ € 1.195,- Bekijk »
<input type="checkbox"/> Vergelijk Samsung QLED QE55Q8D Zwart ★★★★★ € 1.051,- Bekijk »	<input type="checkbox"/> Vergelijk Samsung QLED QE55Q80RAL Zilver, Zwart ★★★★★ € 1.129,- Bekijk »	<input type="checkbox"/> Vergelijk Sony Bravia KD-55XG9505 Zwart ★★★★★ € 1.195,- Bekijk »	<input type="checkbox"/> Vergelijk Sony KD-55AF8 Zwart ★★★★★ € 1.395,- Bekijk »

(Tweakers, 2020b)

Do these indicators contribute to the credibility of the recommender system of Tweakers? (In the information table, the indicators are aligned with the types of bias and overspecialisation. The indicators are presented in bold italics. With 'ease of use', you need to check if the recommender system was easy to use or not.

	Yes	Maybe	No
Complete	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Representative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Consistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Concise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Well-presented	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Objective	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No spin manipulation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Professional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Will have impact	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fair	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ease of use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

After checking the indicators, please rate the credibility of the recommender system of Tweakers again (Scale: 1: very poorly to 7: very well).

Reminder: Accuracy: Are the recommendations similar to the main product and relevant for you? Authenticity: Do you trust the recommendations? Believability: Do you believe that the recommendations are the ones you need?

	1	2	3	4	5	6	7
Accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Authenticity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Believability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What is your gender

- ☐ Male
- ☐ Female

What is your age?

What is your current or highest completed degree?

- ☐ None
- ☐ Less than high school graduate
- ☐ High school
- ☐ Associate degree (MBO)
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Doctorate
- ☐ Other

What is your nationality?

▼ Afghanistan ... Zimbabwe

EMAIL OPTIONAL: If you want to receive the results of my research or other information regarding to my thesis, please type your e-mail address down below

Dutch version BOIT and three-item credibility scale

Vooroordelen en manipulatie	Uitleg
Beoordelingsvooordeel	<i>Slecht gepresenteerd</i> met cijfermatige beoordelingen en <i>onprofessioneel</i> (niet gebaseerd op studies of onderzoek).
Seriële positie-effecten	<i>Inconsistente</i> lijst met vooraf bepaalde producten aan het begin en het eind van de lijst.
Lokeffecten	<i>Incomplete</i> lijst met alleen vooraf bepaalde producten zonder concurreerde producten. Dit is niet <i>objectief</i> en niet <i>representatief</i> .
Risico-aversie	Minder risicovolle items die minder nuttig zijn voor de consument. Dit is <i>spinmanipulatie</i> en maakt AS minder <i>beknopt</i> .
Populariteitsvooordeel	<i>Incomplete</i> lijst met alleen populaire en bekende producten. <i>Deskundige</i> AS voegen ook minder populaire producten van andere merken toe.
Onverwachtheid	Onverwachte en onbekende producten die nuttig zijn <i>hebben een impact</i> op het gedrag van de consument.
Filterbubbel en echokamer	Wanneer consumenten een <i>oneerlijke</i> lijst krijgen met producten van alleen 1 of 2 merken met versterkte overtuigingen.

	Ja	Misschien	Nee
Compleet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Representatief	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Consistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Beknopt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Goed gepresenteerd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Objectief	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Geen spinmanipulatie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Deskundig	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Professioneel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Zal impact hebben	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Eerlijk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Eenvoudig te gebruiken	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Herinnering: Nauwkeurigheid: Zijn de aanbevelingen vergelijkbaar met het hoofdproduct en relevant voor u? Juistheid: Vertrouwt u de aanbevelingen? En aannemelijkheid: Gelooft u dat de aanbevelingen degene zijn die u nodig hebt?)

	1	2	3	4	5	6	7
Nauwkeurigheid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Juistheid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Aannemelijkheid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>