Accounting for Heuristics in Reputation Systems: An Interdisciplinary Approach on Aggregation Processes

Dirk van Straaten[a,*], Vitalik Melnikov[b], Eyke Hüllermeier[c], Behnud Mir Djawadi[a], René Fahr[a,d]

[a] *Paderborn University, Department of Management, Heinz Nixdorf Institute, Warburger Str. 100, 33098 Paderborn, Germany*
[b] *Paderborn University, Department of Computer Science, Heinz Nixdorf Institute, Warburger Str. 100, 33098 Paderborn, Germany*
[c] *Ludwig-Maximilians-University Munich, Institute for Computer Science, Oettingenstr. 67, 80538 Munich, Germany*
[d] *Institute for the Study of Labor (IZA), Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany*

Abstract

Aggregation metrics in reputation systems are important for overcoming information overload. When using these metrics, technical aggregation functions such as the arithmetic mean are implemented to measure the valence of product ratings. However, it is unclear whether the implemented aggregation functions match the inherent aggregation patterns of customers. In our experiment, we elicit customers' aggregation heuristics and contrast these with reference functions. Our findings indicate that, overall, the arithmetic mean performs best in comparison with other aggregation functions. However, our analysis on an individual level reveals heterogeneous aggregation patterns. Major clusters exhibit a binary bias (i.e., an over-weighting of moderate ratings and under-weighting of extreme ratings) in combination with the arithmetic mean. Minor clusters focus on 1-star ratings or negative (i.e., 1-star and 2-star) ratings. Thereby, inherent aggregation patterns are neither affected by variation of provided information nor by individual characteristics such as experience, risk attitudes, or demographics.

Keywords: customer reviews, aggregation, heuristics, binary bias, arithmetic mean

JEL Classification Numbers: D81, D12, C91

* Corresponding author
Email addresses: dirk.van.straaten@upb.de (Dirk van Straaten), melnikov@mail.upb.de (Vitalik Melnikov), eyke@ifi.lmu.de (Eyke Hüllermeier), behnud.mir.djawadi@upb.de (Behnud Mir Djawadi), rene.fahr@upb.de (René Fahr)

1. Introduction

In online shopping, there are many sources of information available for the customer to build up the purchase decision on. One of the most important sources is customer feedback since it is the only information not provided by the manufacturer or selling platform. Hence, it is not surprising that over 80% of the customers use them during the purchase process (Cheung and Lee 2012). As there are plenty of customer reviews it is difficult for customers to integrate all of them in an non-aggregated manner into the decision process. Hence, selling platforms provide aggregated measurements in which the numerical part of customer reviews, the ratings, are processed to single index values representing the valence (i.e., the quality) of the underlying product. In practice, the arithmetic mean is often employed to calculate the valence of customer ratings. Thereby, the star categories are weighted in accordance to their scale values. However, literature in the field of psychology and behavioral economics identify plenty of behavioral biases that can, for instance, be driven by bounded rationality or employed heuristics (cf. Tversky and Kahneman 1974; Gigerenzer and Todd 1999). In this paper, we investigate to which degree inherent heuristics of customers have an effect on the aggregation of customer rating distributions and whether they result in systematic biases that should be addressed in the implemented aggregation metrics in reputation systems.

Therefore, we develop an experimental design in which subjects receive triples of customer rating distributions and are asked to rank these in accordance with their preferences. Ensuring to elicit real preferences, we implement incentives as customer rating distributions are partially linked to real products from an online marketplace. Subjects have a higher chance to win the underlying products when they rank a product better. We analyze these ranking decisions by employing a Maximum-Likelihood approach. In particular, we estimate the category weights (i.e., 1-star, ..., 5-star) with Plackett-Luce model specifications for each subject and compare these estimates with weights of the arithmetic mean and other aggregation functions that, for instance, correspond to minimizing 1-star ratings.

The results confirm the arithmetic mean to be the best predictor of behavior in comparison with other aggregation functions. However, our cluster analysis reveals that the majority of subjects also exhibit the binary bias. That is, moderate categories (i.e., 2-star and 4-star) are over-weighted and extreme categories (i.e., 1-star and 5-star) are under-weighted. As minor clusters also show other aggregation patterns by only minimizing 1-star, or negative (i.e., 1-star and 2-star) ratings, we identify heterogeneity in aggregation behavior. Contrary to our predictions, aggregation patterns are not affected by demographics, risk attitudes, or experience in online shopping. Moreover, these patterns are robust, independent of whether additional numerical information is provided or not.

Providing novel insights into customers' aggregation patterns, our study has important implications: Online marketplaces should consider the binary bias in aggregation metrics to measure the valence of product ratings. Taking into account the heterogeneity in aggregation patterns of customers, marketplaces could

also implement instruments to elicit customers' aggregation preferences and provide individual aggregation metrics for the valence of customer ratings. Therefore, our experimental design can serve as a role model.

The subsequent sections are organized as follows. In Chapter 2, we explain customer rating distributions, aggregation functions, and aggregation heuristics. We survey the literature in this field and derive propositions with regard to the aggregation heuristics of the customers. In Chapter 3, we explain the experimental design to test our propositions. After describing the structure of the data and our analysis approach in Chapter 4, we describe our results in Chapter 5. Chapter 6 concludes the article.

2. Aggregation Processes in Reputation Systems

Online purchase decisions are decisions under uncertainty as the outcome is uncertain, in particular, the quality and satisfaction derived from experiencing the purchased product. Customer reviews can reduce the uncertainty as previous customers share their experience and, thereby, substitute their own lacking experience. Dellarocas (2005) calls this challenge in online markets *adverse selection* and shows theoretically that reputation systems are the remedy for this problem. This result is supported by Chevalier and Mayzlin (2006) who show that a good reputation positively affects sales. Meta analyses (Floyd et al. 2014; You et al. 2015) find robust effects of customer ratings on sales resp. volume and valence elasticities.[1] Thereby, customer reviews consist of a numerical evaluation of the customer's experience with the rated product or service, the customer rating, and a free text enabling the customer to go more into detail and justify his or her rating. The scale of the numerical evaluation differs from marketplace to marketplace. Examples are a dichotomous scale (positive, negative) or star ratings (range from one to five stars). As we are interested in the aggregation processes of the quantitative information in reputation systems, we focus on numerical evaluations, i.e., customer ratings (cf. Figure 1).

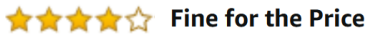


Figure 1: Example of 4-star rating from an Amazon.com website.

2.1. Customer Ratings in Reputation Systems

Motives of reviewers. There are manifold motives for providing customer reviews in reputation systems. There is empirical evidence for altruism (cf. Peddibhotla and Subramani 2007; Hennig-Thurau et al. 2004; Munzel and H. Kunz 2014; Cheung and Lee 2012; Wu 2019; Hoyer and van Straaten 2021; van Straaten 2021), reciprocity towards peers (cf. Peddibhotla and Subramani 2007; Hennig-Thurau et al. 2004; Munzel and H. Kunz 2014) and sellers (cf. Munzel and H. Kunz 2014; Wu 2019; van Straaten 2021), economic

[1] Considering endogeneity of reviews and sales, the results in Duan et al. (2008) indicate, however, that the valence of ratings does not impact the sales.

incentives (cf. Peddibhotla and Subramani 2007; Hennig-Thurau et al. 2004; Wu 2019; van Straaten 2021), self-expression, building up a reputation, and social affiliation to drive reviewing behavior (cf. Peddibhotla and Subramani 2007; Hennig-Thurau et al. 2004; Munzel and H. Kunz 2014; Cheung and Lee 2012; Wu 2019; Hoyer and van Straaten 2021). Which motives are addressed depends on the design of the reputation system (Gutt et al. 2019).

Stimulation of customer reviews. The results of Askalidis et al. (2017) show that email invitations to publish a review are a valid remedy to decrease the selection bias by collecting more reviews of otherwise not reporting customer groups. This approach is implemented in a way that also prevents the occurrence of a *social influence bias* (cf. Sundar et al. 2008), i.e., the unconscious effect of read reviews on one's own reviewing behavior. However, incentivized reviews are perceived less helpful in comparison with self-motivated reviews. There are other approaches for stimulating the provision of customer reviews by addressing the motives mentioned above (cf. Marinescu et al. 2018 or van Straaten 2021 for an overview).

Representativeness of customer ratings. These stimulations are necessary as literature shows that customer ratings are biased and, hence, not normally distributed. This goes along with distorted arithmetic mean values that do not reflect true product quality (Hu et al. 2006, 2009). Hu et al. (2009) identifies two biases that can explain the bimodal histogram of customer ratings for products (called the j-shape). The *purchasing bias* explains the majority of ratings in the best category by considering the effort of search information and consciously selecting the product before the purchase. This increases the likelihood to be fully satisfied with the purchased product. They also introduce the *under-reporting bias* of moderately satisfied customers. Without being very satisfied or dissatisfied with the product, the intrinsic motivation is not sufficient to exceed the effort of publishing a review. The j-shape is identified on many online platforms (cf. Schoenmueller et al. 2020). Although customers are aware of these (self-selection) biases in customer ratings, they cannot infer the true quality due to bounded rationality (Hu et al. 2017). In accordance with the under-reporting bias, Ho et al. (2017) focus on the *disconfirmation bias*, i.e., the gap between pre-purchase expectations and the post-purchase experience. Results indicate a higher propensity of customers to publish a rating when the magnitude of disconformity is larger. Ratings that are motivated by disconformity are also biased in the direction of deviation from expectation. Pointing to the vulnerability of reputation systems due to self-selection biases, Li and Hitt (2008) show that idiosyncratic preferences of early reviewers can strongly bias aggregation measures.[2]

2.2. Aggregation Metrics

Aggregation metrics are used to describe the ratings with parameters by summarizing customer ratings (cf. Figure 2). Dellarocas (2003) show challenges with regard to electronic word-of-mouth and address

[2] They identified an undershooting of 0.16 star points, which corresponds to one forth of standard deviation across all products examined.

Figure 2: Example of customer rating distribution and calculated valence from an Amazon.com website.

the aggregation of customer reviews as an important topic to provide accurate information about sellers and products. For instance, Camilleri (2017) illustrates that outliers are discounted in customer reviews when they are presented in disaggregated form because these are attributed to reviewer (instead of product) reasons. In aggregated form, i.e., the distribution is shown but not the single ratings, this effect disappears. Wulff et al. (2015) provide evidence for the need of visualization and summaries by showing in a hypothetical setting that participants, when faced with distributions or a sample of single ratings, underestimate in the sample setting the occurrence of rare events, leading more often to a decision for a product with a lower objective mean. Thereby, dependencies between customer reviews and aggregation metrics are bidirectional. That is, the rating scores of products are not only calculated by the single ratings, but also affects the credibility of customer reviews that are in conflict with the overall evaluation (Qiu et al. 2012).

The purpose of aggregation metrics is to achieve comparability between products on marketplaces and to evaluate product quality. Besides the *valence* of reviews (i.e., a value of quality or goodness), the number or ratings (*volume*) and the *variance* of the ratings also play a role. You et al. (2015) show that valence and volume positively affect the sales elasticity. Zimmermann et al. (2018) point out that variance in customer ratings caused by quality differences has a negative effect on the demand. Thereby, variance or bimodality are not necessarily a negative component of customer rating distributions. Rozenkrants et al. (2017) find that in dimensions reflecting self-expressing motives (such as style) bimodal distributions are partially preferred. Similar results are reported by He and Bond (2015).

There are studies investigating interdependencies between metrics. In Watson et al. (2018) the impact of the volume of reviews under different average ratings is investigated. As a key result they find a preference shift from products with a higher average and lower number of reviews to products with a lower average but higher numbers of reviews. Coba et al. (2019) show that customers are more sensitive to the mean and number of ratings compared to the variance or origin of the rating. Flanagin et al. (2011) show that the average rating is important in assessing the quality of the underlying product. However, the volume of the ratings does not have a significant impact on this evaluation.

There are many possible metrics to determine the valence of products by means of aggregated customer

ratings. Overviews over aggregation functions and their attributes are provided comprehensively in Beliakov et al. (2011) and Garcin et al. (2009). Focusing on the arithmetic mean as the reference aggregation function, we also provide evidence for aggregation metrics with behavioral foundations, i.e., identified heuristics in psychology. Subsequently, we consider customer rating distributions that contain categories $k = 1, ..., 5$ and the relative frequency x_k of category k (cf. Figure 3). Valence v is a function of category weights W_k and relative frequencies x_k:

$$v = \sum_{k=1}^{5} W_k \times x_k \tag{1}$$

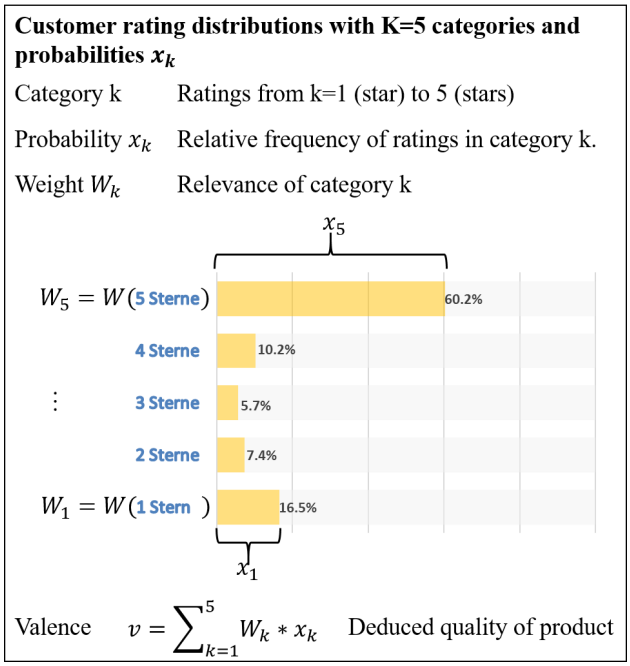


Figure 3: Measuring valence of customer rating distributions with category weights W_k and probabilities x_k.

2.2.1. Measuring Valence with the Arithmetic Mean

The most prominent approach to measure the valence of customer rating distributions resp. products is to employ the arithmetic mean. Thereby, each category is weighted in accordance with its ordinal category scale, i.e., $W_k = k$:

$$v_{AM} = \sum_{k=1}^{5} k \times x_k.$$

In a meta-analysis it was identified that rating valence does not affect perceived helpfulness (Hong et al. 2017). That is, there is no evidence that positive or negative reviews are weighted differently. Hence, considering all ratings equally and weighting them in accordance with their valuation is equal to employing the arithmetic mean. Hurley and Estelami (1998) showed that the arithmetic mean is the best predictor for

quality in comparison with other different metrics such as median, mode, kurtosis, or skewness. Comparing consumer reports (i.e., objective quality assessment) with customer ratings, the findings of De Langhe et al. (2015) show a low correlation indicating a rather bad cue for product quality. However, the mean of ratings affects customers stronger than price or the volume of ratings, raising evidence for the appropriateness of the arithmetic mean as a measure of valence. McGlohon et al. (2010) investigate the explanatory power of different aggregation metrics. Their results indicate that the arithmetic mean performs well and similarly compared to more sophisticated algorithms. Given this evidence, we propose that the arithmetic mean is indeed a good predictor and a metric that is inherently employed by subjects to deduce the quality of products.

Proposition 1. *Provided with product rating distributions, subjects inherently apply the arithmetic mean to deduce the quality of products.*

2.2.2. Alternative Measures of Valence

We propose that the arithmetic mean is applied to measure the valence of product rating distributions. However, considering biases from cognitive science and behavioral economics we discuss alternatives to the arithmetic mean that might be more congruent with the customers' assessment of the valence of customer rating histograms. We phrase these effects *heuristics*, following the definition of Tam and Ho (2005) that heuristic information processing means that "people consider a few informational cues – or even a single informational cue – and form a judgment based on these cues." Heuristics are highly economical and usually effective, but also produce systematic errors (Tversky and Kahneman 1974). Gigerenzer and Todd (1999) introduce fast and frugal heuristics that are applied to limit searches through stopping rules, e.g., when deducing the valence of products. Thereby, heuristics can be employed consciously as well as unconsciously (Mousavi and Gigerenzer 2014). Zhang et al. (2014) consider systematic and heuristic factors in the analysis of customer reviews and their effect on purchase intention. Thereby, heuristic cues (i.e., perceived quantity of reviews and perceived source credibility) affect behavioral intention directly and also affect the perception of systematic argument quality.

Focus on the highest category (FIV) / Follow the majority. The endorsement heuristic causes customers to focus on the majority of subjects (cf. Metzger et al. 2010). In most cases this means to focus on five star reviews.[3] Considering only the highest category, i.e., 5-star ratings, indicates focusing only on the best experiences. Here, a customer rating distribution is assessed only by the relative frequency of the 5-star ratings and other categories are not taken into account, i.e.: W_1^{FIV}, ..., $W_4^{FIV} = 0$, $W_5^{FIV} = 1$.

$$v_{FIV} = \sum_{k=1}^{5} W_k^{FIV} \times x_k = x_5$$

[3] Note that without the j-shape of customer ratings (Hu et al. 2009) this heuristic would correspond to the mode function.

Focus on the lowest category (ONE). Lee et al. (2009) show that extremely negative ratings have a stronger impact on consumers in comparison with moderately negative ratings. Hence, focusing on 1-star ratings is also a reasonable heuristic for evaluating customer rating distributions. Considering only the lowest category, i.e., 1-star ratings, indicates focusing only on the worst experiences. Here, a customer rating distribution is assessed only by the relative frequency of the negative 1-star ratings, i.e.: $W_1^{ONE} = -1;\ W_2^{ONE},\ ...,\ W_5^{ONE} = 0.$

$$v_{ONE} = \sum_{k=1}^{5} W_k^{ONE} \times x_k = -x_1$$

Binary perception of ratings (BIN). Fisher et al. (2018) investigate the influence of customer rating distributions and the displaying of the mean on hypothetical purchase decisions. They identify the binary bias: i.e., participants see 4-star and 5-star ratings only as positive and 1-star and 2-star ratings as negative ratings and do not sufficiently discriminate between categories. This corresponds to the following weights: $W_1^{BIN} = W_2^{BIN} = -1,\ W_3^{BIN} = 0,\ W_4^{BIN} = W_5^{BIN} = 1.$

$$v_{BIN} = \sum_{k=1}^{5} W_k^{BIN} \times x_k = -x_1 - x_2 + x_4 + x_5$$

Focus on positive ratings (POS) / positive valence. Sundar et al. (2009) identify the bandwagon heuristic in the processing of customer ratings, i.e., customers want to be on the winner's side and choose the product that has the most positive ratings. Cognitive science identified that the majority of people overweight the occurrence of positive states in the future due to a superiority illusion and unrealistic optimism (cf. Sharot and Garrett 2016). Transferring this general tendency to the domain of customer ratings means focusing the positive customer reviews and ignoring the occurrence of negative experiences. Combining this focus on positive events with the binary perception of customer ratings (cf. Fisher et al. 2018), another heuristic might exist by taking the relative frequencies of 4-star and 5-star ratings (POS) as decisive criteria. Indicating that customers focus on moderate chances, this strategy means to value the customer rating distributions according to the sum of the relative frequencies of the 4- and 5-star ratings and ignore the 1- to 3-star ratings, i.e.: $W_1^{POS},\ ...,\ W_3^{POS} = 0;\ W_4^{POS} = W_5^{POS} = 1.$

$$v_{POS} = \sum_{k=1}^{5} W_k^{POS} \times x_k = x_4 + x_5$$

Focus on negative ratings (NEG) / negative valence. In contrast, there is also evidence in Bae and Koo (2018) that customers focus more on negative reviews (e.g., due to skepticism about fake reviews (Luca and Zervas 2016)). In the domain-specific context of online purchases and the dealing with plenty of customer ratings, Chen et al. (2018) investigate the change from one- to multidimensional reputation systems and thereby find that the aggregation of sub-dimensions to the overall evaluation is not processed by weighting each dimension equally and forming the average. As multidimensional reputation systems show a higher

average rating, this means that reviewers show the tendency to focus on negative aspects. Thereby, Fisher et al. (2018) provide evidence that there is not sufficient discrimination between 1-star and 2-star ratings, thus resulting in the binary bias. When negative ratings, i.e., 1-star and 2-star ratings (NEG), are decisive this indicates the customers' focus on the moderate risks. In this case, the assessment of a customer rating distribution is determined only by the negative sum of the 1- and 2-star ratings and the relative frequencies of 3- to 5-star ratings are not considered, i.e.: $W_1^{NEG} = W_2^{NEG} = -1;\ W_3^{NEG},\ ...,\ W_5^{NEG} = 0$.

$$v_{NEG} = \sum_{k=1}^{5} W_k^{NEG} \times x_k = -(x_1 + x_2)$$

Median (MED) - control for outliers. Another approach to reduce biases by outliers is introduced in Garcin et al. (2013). They suggest using the truncated arithmetic mean, where the α smallest and α largest reported ratings are truncated before the arithmetic mean is calculated. However, the median elicit the most truthful ratings. More precisely, Yaniv (1997) finds that weighting and trimming are two employed heuristics in situations where subjects are confronted with questions about unknown facts. When answers of other subjects to these questions are provided, the extreme answers are trimmed and the moderate answers are weighted (e.g., more weight on the modest answers). Transferred on customer ratings, extreme values such as 1-star or 5-star ratings might be trimmed and the modest ratings are weighted in accordance with personal preferences. Considering the approach to control for outliers, the median is a reasonable approach, as it is theoretically the most moderate and robust metric (Garcin et al. 2009). It means assessing a customer rating distribution in accordance with the value separating the higher half from the lower half of the customer ratings. Findings of Hurley and Estelami (1998) suggest that the median performs better than other aggregation functions such as mode, skewness, or kurtosis (although worse than the arithmetic mean). Jurca et al. (2010) also argue in favor of the median to aggregate ratings, thereby addressing the skewness of distributions.

Considering the aforementioned biases and alternative aggregation functions, we propose heterogeneity in aggregation heuristics.

Proposition 2. *Provided with product rating distributions, subjects apply heterogeneous aggregation heuristics to deduce the quality of products.*

2.3. Robustness of Heuristics and Determinants

Another goal of this study is to examine the robustness of heuristics applied by customers in their evaluation of customer rating distributions. Hauser (2014) presents various decision heuristics in consumer theory that are part of the adaptive toolbox (Gigerenzer and Todd 1999): that is, heuristics are adaptable and context-specific. Hauser (2014, p.1690) argues that the applied heuristics are often robust to missing data, which means that the variation of the provided information should not result in different behavior.

However, Reyes et al. (1980) give evidence for the availability bias that gives more weight to easily available information in decision processes. When numerical information is available and easier to interpret, this could result in different decisions. Moreover, Metzger and Flanagin (2013, p.216) describe the expectancy violation heuristic, i.e., judging the credibility is affected by the presentation of information on websites. When subjects expect quantitative information but only see visual information, they might change evaluations and hence behave differently. Adomavicius et al. (2019) identify that the valence of ratings depends on the visual summary presentations of ratings in recommendation systems before consumption. When this information is presented numerically, the ratings after consumption are more negative: i.e., numbers seem to have a stronger impact. This is in contrast with Kostyk et al. (2017) who analyze the impact of the display condition on purchase intention. When only a single bar with the mean value is displayed, the purchase intention is higher in comparison with decisions in which rating distributions (with the same mean values) are displayed. Thus, the aggregate value seems to be more important in purchase decisions.

In conclusion, empirical evidence of the effect of the amount of information on heuristic processes is ambiguous. However, literature in the context of customer ratings observe changes in behavior. Hence, we propose:

Proposition 3. *Subjects' applied heuristics to deduce the quality of products are affected by variations of provided information.*

Since we observe individual decision making based on behavioral patterns and heuristics, we investigate the impact of *individual characteristics*. The results of Cheung and Thadani (2012) reveal an impact of the individual's characteristics on the perception of stimulus sent in eWOM communication. In particular, Yin et al. (2018) identify that heuristics for credibility evaluation differ between men and women. Besedeš et al. (2012) show that the employment of heuristics also depend on age, as older subjects show a higher propensity to choose suboptimal heuristics.

Especially, *experience in online shopping* might be a decisive factor when being faced with aggregation decisions in the context of customer ratings. That is, the accuracy of given customer ratings might be perceived differently from experts in online shopping and customers who only buy a product once a year online. Metzger and Flanagin (2013, p. 217) also state that users with more experience employ different heuristics. In von Helversen et al. (2018) older and younger adults faced hypothetical online purchasing decisions. Younger participants claimed to be more experienced and in most cases, contrary to the older adults, decided in favor for the product with the higher average ratings.

Finally, Wu and Chang (2007) show that *risk attitudes* are correlated with online shopping behavior. Metzger and Flanagin (2013, p. 215) describe the self-confirmation bias, highlighting the tendency to focus on the information that is consistent with the own expectations. Risk-averse subjects might focus more on bad outcomes and hence show different behavior in the evaluation of product rating distributions. Hence,

we propose:

Proposition 4. *Subjects' applied heuristics to deduce the quality of products are affected by individual characteristics, experience, and risk attitudes.*

3. Experimental Design

In comparison with most of previous studies, which derive an optimal ranking and compare it with the actual behavior, we let the subjects rank various customer rating distributions and infer the participants' aggregation heuristics.[4] This allows us to find the optimal aggregation function and compare it with different above-mentioned aggregation functions.

The implementation of the experiment is as follows. Subjects receive customer ratings of three products and are asked to rank the products according to their preferences. The customer ratings differ with regard to their relative frequencies and their arithmetic means. An example of these aggregated customer ratings is depicted in Figure 4.

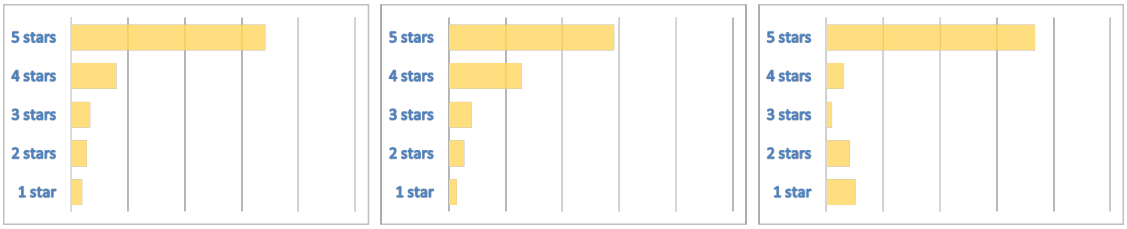


Figure 4: Illustrative bundle of three aggregated customer ratings.

Participants only see the products' aggregated customer ratings, but do not know the products' name or detailed specifications except that the aggregated customer ratings are from the same product category and have similar prices and specifications. We thereby use aggregated customer ratings that allow us to separate rankings on the basis of different decision heuristics, such as minimization of negative ratings or maximization of positive ratings from rankings that favor the arithmetic mean.

Subjects choose rankings for overall 12 categories, whereby the distributions to rank in these categories are partially artificial. Six decisions are based on real aggregated customer ratings from the Amazon marketplace. Thereof, three decisions are used to incentivize the subjects' decisions. In particular, subjects receive the USB flash drive they rank first or second as payment and, in addition, are given the chance to win another product they choose from one of two other product categories, in particular a tablet computer or a gooseneck (i.e., tablet holder). Thereby, participants have a 70% chance of winning the product they rank first, and a 30% chance for the product they rank second. Thus, the complete ranking decision is incentivized. The

[4] This approach is similar to Yang et al. (2016) who use comparative data between products to derive ranking of products.

decisions are ordered randomly in a way that participants do not know which decision determines their payoff. Using artificial aggregated customer ratings allow us to investigate the employed decision heuristics more precisely. Especially, we can disentangle heuristics on the basis of the arithmetic mean and the median. The used aggregated customer ratings are shown in Table A.11.

We employ two treatments. In the *control treatment* (CT), subjects are only given the aggregated customer ratings without any additional information (cf. Figure 4). In the *information treatment* (IT), subjects additionally see the relative frequency of each of the star categories and the value of the arithmetic mean associated to the distribution (cf. Figure 5). The treatment variation enables investigating the impact of information's degree on aggregation heuristics (Proposition 3).

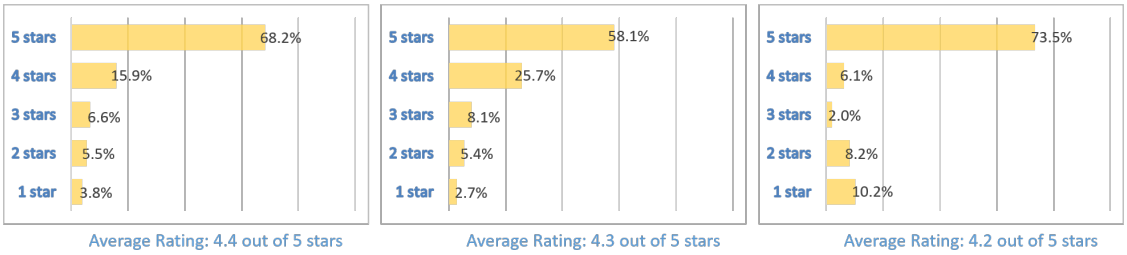


Figure 5: Illustrative bundle of three aggregated customer ratings in the information treatment. Relative frequencies and the arithmetic mean are provided.

To elicit further heuristics, we conduct a questionnaire after the experiment, asking the subjects about their decision criteria. Additionally, we survey socio-demographic characteristics including risk preferences and experience in online shopping to investigate Proposition 4.

Using the online recruiting system ORSEE (Greiner 2015), participants were recruited from a pool of approximately 2,800 students from different fields of study who volunteered to become prospective participants in economic experiments. We conducted our experiment at the Business and Economic Research Laboratory (BaER-Lab) at Paderborn University, Germany, in December 2018. We conducted four sessions with a total of 107 participants. The experiment was computerized and conducted using the software z-Tree (Fischbacher 2007). In each session, participants received the same introductory talk and were told not to communicate with each other for the duration of the experiment. Participants read the written instructions and could ask questions individually and in private to the experimenter. Afterwards, the experiment was started. Sessions lasted 75 minutes on average and participants earned material prizes with average values of EUR 19.10. Additionally, each participant earned a show-up fee of EUR 2.50.

4. Data and Statistical Model

The data gathered from our laboratory experiment can essentially be summarized in the form of a matrix (separately for the control and the information treatment) as follows:

$$Z = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,m} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n,1} & z_{n,2} & \cdots & z_{n,m} \end{pmatrix}$$

Here, the rows correspond to the participants (53 for CT and 54 for IT), and the columns to object groups (12 in total, same groups for both CT and IT). Each element $z_{i,j}$ in either of the above matrices holds the preference of a participant i over a product group j, that is, the ranking (total order)

$$z_{i,j} : \boldsymbol{o}_{\pi_{i,j}(1)} \succ \boldsymbol{o}_{\pi_{i,j}(2)} \succ \boldsymbol{o}_{\pi_{i,j}(3)} \tag{2}$$

of three different objects. Here, $\pi_{i,j}$ is a permutation $\{1,2,3\} \longrightarrow \{1,2,3\}$ such that $\pi_{i,j}(k)$ is the index of the object on position k in the ranking. The objects themselves are partially real products from the amazon.de[5] marketplace. Every product $\boldsymbol{o}_i$ is represented by the customer rating distribution

$$f(\boldsymbol{o}_i) = (x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}) \in [0,1]^5\,, \tag{3}$$

where $x_{i,k}$ is the relative frequency of k-star ratings of the product. These ratings are provided in Table A.11.

To analyze and test our propositions empirically, a suitable stochastic model of the data-generating process is required. According to our assumptions, the latter consists of two stages. Being confronted with a set of three choice alternatives, a participant first evaluates each alternative in terms of a (latent) utility degree, and then sorts them in decreasing order of preference according to these degrees. To account for inaccuracies, mistakes, and other random effects, the stochastic nature of the model is clearly important.

4.1. Plackett-Luce Model

Since the observational data consists of rankings (2), we use the so-called Plackett-Luce (PL) model (Plackett (1975); Luce (1959)), which is a model of rank data that is parametrized by quantitative preference degrees for individual choice alternatives, and hence nicely complies with the assumptions of our data-generating process.

More specifically, the PL model defines a probability distribution of the set of all rankings of a given set $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_K\}$ of K choice alternatives, that is, on the set of all permutations of $[K] = \{1, \ldots, K\}$. It is parametrized by a vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_K) \in \mathbb{R}_+^K$, where each v_i can be interpreted as the weight

[5] www.amazon.de

or "strength" of the option $\boldsymbol{o}_i$. The probability assigned by the PL model to a ranking represented by a permutation $\pi \in \mathbb{S}_K$ is given by

$$p_{\boldsymbol{v}}(\pi) = \prod_{i=1}^{K} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \ldots + v_{\pi(K)}}. \tag{4}$$

The product on the right-hand side of (4) is the probability of producing the ranking π in a *stagewise* process. First, the item on the first position is selected, then the item on the second position, and so forth. In each step, the probability of an item to be chosen next is proportional to its weight. Consequently, items with a higher weight tend to occupy higher positions. In particular, the most probable ranking (i.e., the mode of the PL distribution) is simply obtained by sorting the items in decreasing order of their weight:

$$\pi^* = \operatorname*{argmax}_{\pi \in \mathbb{S}_K} p_{\boldsymbol{v}}(\pi) = \operatorname*{argsort}_{k \in [K]} \{v_1, \ldots, v_K\}. \tag{5}$$

In our case, $K = 3$, v_j represents the latent utility of the j^{th} product. For example, $\pi = (\pi(1), \pi(2), \pi(3)) = (2, 3, 1)$ represents the ranking $\boldsymbol{o}_2 \succ \boldsymbol{o}_3 \succ \boldsymbol{o}_1$, according to which the second product is the most preferred one, the third product the second best, and the first product the least preferred. The PL probability to observe this ranking is

$$p_{\boldsymbol{v}}(\pi) = \frac{v_2}{v_1 + v_2 + v_3} \times \frac{v_3}{v_1 + v_3} \times \frac{v_1}{v_1}\,.$$

4.2. Product Preferences

In a general case we assume each product $\boldsymbol{o}_j$ to be characterized in terms of descriptive statistics (or features) $(x_{j,1}, \ldots, x_{j,k})$ of the underlying customer rating distribution. It appears reasonable to model the utility v_j as an aggregation of these features:

$$v_j = A\big(x_{j,1}, \ldots, x_{j,n}\big)\,,$$

where A is a suitable aggregation function. Concretely, we assume v_j to be a log-linear function of a *generalized mean*:

$$v_i = \exp\left(\alpha \sum_{k=1}^{5} x_k\, w_k\right) = \exp\left(\alpha \langle \boldsymbol{x}, \boldsymbol{w} \rangle\right), \tag{6}$$

where the coefficient w_k reflects the importance of the k-star frequency $x_{j,k}$ (cf. Figure 3).

Since PL is invariant with regard to multiplication of the parameter $\boldsymbol{v}$ by a positive constant, and the parameter $\alpha > 0$ accounts for scaling effects, we can normalize the coefficients such that

$$\sum_{k=1}^{5} w_k = 0\,, \quad \sum_{k=1}^{5} |w_k| = 1\,.$$

This allows for a convenient interpretation of the model: The sign of the coefficient w_k determines the direction in which the frequency x_k influences the preference (positive or negative), and the absolute value

the importance of the k-star category relative to the others. The parameter α captures the "precision" of the decision-maker: The larger α, the more probably the produced ranking will agree with the latent utilities. In particular, the probability of the mode (5) converges to 1 for $\alpha \to \infty$. This case corresponds to a perfect decision maker who deterministically ranks in accordance with the latent utilities. The opposite extreme is $\alpha = 0$, which leads to a uniform distribution on the set of rankings. In other words, $\alpha = 0$ corresponds to a decision maker who essentially ignores the utilities and instead sorts the products completely at random.

The model as outlined above restricts the class of aggregation functions A by assuming that the frequency of ratings is combined by means of a weighted average.[6] While this assumption may clearly be questioned, let us note that our class still covers a wide range of important aggregations as special cases, including those mentioned as heuristics by the participants. In particular, the simple arithmetic mean (AM) is recovered as a special case by the following weights:

$$w_1 = -\frac{1}{3},\ w_2 = -\frac{1}{6},\ w_3 = 0,\ w_4 = +\frac{1}{6},\ w_5 = +\frac{1}{3}. \tag{7}$$

Likewise, for example, the "5-star ratio" (FIV) heuristic is recovered by

$$w_1 = -\frac{1}{8},\ w_2 = -\frac{1}{8},\ w_3 = -\frac{1}{8},\ w_4 = -\frac{1}{8},\ w_5 = +\frac{1}{2}.$$

Note that all considered heuristics are covered by the aggregation function (6) as special cases.[7] The only exception being the median heuristic (MED), since it cannot be computed from the provided relative frequencies.

4.3. Parameter Estimation

The model introduced above is parametrized by α and $\boldsymbol{w} = (w_1, \ldots, w_5)$, which capture the precision and the aggregation behavior of a subject, respectively. These parameters determine the PL parameters (6), which in turn determine the probability of rankings (4). Thus, estimating our model comes down to estimating α and $\boldsymbol{w}$. In this section, we tackle this problem using the principle of maximum likelihood estimation.

Suppose that data $\mathcal{D}$ in the form of N rankings $\pi_1, \ldots, \pi_N$ has been observed, where each ranking corresponds to a preferential ordering of three products $\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3$. The likelihood of parameters $\alpha, \boldsymbol{w}$ is then given by the probability of observing this data:

$$\begin{aligned} L(\alpha, \boldsymbol{w}) &= \prod_{n=1}^{N} p_{\alpha,\boldsymbol{w}}(\pi_n) = \prod_{n=1}^{N} \prod_{j=1}^{3} \frac{v_{\pi(j)}}{\sum_{i=j}^{3} v_{\pi(i)}} \\ &= \prod_{n=1}^{N} \prod_{j=1}^{3} \frac{\exp\left(\alpha \sum_{k=1}^{5} w_k\, x_{\pi(j),k}\right)}{\sum_{i=j}^{3} \exp\left(\alpha \sum_{k=1}^{5} w_k\, x_{\pi(i),k}\right)} \end{aligned} \tag{8}$$

[6] See, e.g., Josang et al. (2007) for other classes of aggregation measures.
[7] Table A.12 provides theoretical and estimated weights.

where $\boldsymbol{x}_j = (x_{j,1}, \ldots, x_{j,5})$ is the frequency distribution of the j^{th} product, and $p_{\alpha,\boldsymbol{w}}$ the PL probability with parametrization (6). Thus, the ML estimate is obtained as

$$(\alpha^*, \boldsymbol{w}^*) = \arg\max_{\alpha,\boldsymbol{w}} \sum_{n=1}^{N} \sum_{j=1}^{3} \Bigg(\alpha \sum_{k=1}^{5} w_k \, x_{\pi(j),k} - \log\Bigg(\sum_{i=j}^{3} \exp\Bigg(\alpha \sum_{k=1}^{5} w_k \, x_{\pi(i),k} \Bigg) \Bigg) \Bigg) .$$

To show the convexity of the negative log-likelihood function, or equivalently the concavity of (8), we reparametrize the model (6) as follows:

$$\begin{aligned} v_i = \exp\left(\alpha \sum_{k=1}^{5} x_k \, w_k \right) = \exp\left(\sum_{k=1}^{5} x_k \, \alpha w_k \right) \\ = \exp\left(\langle \boldsymbol{x}, \alpha \boldsymbol{w} \rangle \right) = \exp\left(\langle \boldsymbol{x}, \boldsymbol{w'} \rangle \right) , \end{aligned} \tag{9}$$

with

$$\sum_{k=1}^{5} w'_k = 0 \, , \quad \sum_{k=1}^{5} |w'_k| = \alpha \, .$$

The resulting model is well-known in the preference learning literature as the *Plackett-Luce model with features* (Cheng et al. (2010)). It was already shown by Schäfer and Hüllermeier (2018) that the negative log-likelihood of this model is convex. The authors also prove that the more general (bilinear) Plackett-Luce model is identifiable which implies the identifiability of our model.

Therefore, the parameter estimation can be accomplished using quasi-Newton type algorithms such as L-BFGS-B (Byrd et al. (1995)). One technical problem may occur in the (unlikely) case where a parameter $\boldsymbol{w}$ exists such that all rankings $\pi_1, \ldots, \pi_N$ are in perfect agreement with the (unscaled) utilities $\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle$, i.e., where the objects $\boldsymbol{o}_j$ are always sorted in decreasing order of the values $\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle$. In this case, the likelihood function can be made arbitrarily large by increasing α, i.e., we would estimate $\alpha^* = \infty$. To avoid this problem, we put an upper bound on α^*.

In our setting, parameters can be estimated on the basis of different data sets $\mathcal{D}$. We will consider a *subject-wise* setting, where a separate model is fitted for every subject. Thus, preferences are allowed to change between subjects, but are assumed to be constant over all decisions within a subject.

5. Results

Demographics of the subjects are provided in Table 1.

Table 1: Demographic information of participants in information and control treatment.

		IT (n=54)	CT (n=53)	Total
Male		37%	32%	35%
Age		22.1 (2.9)	22.2 (2.7)	22.1 (2.8)
Semester		3.3 (3.2)	4.5 (3.2)	3.9 (3.2)
Studies:	Economics	33.3%	39.6%	36.5%
	Education	42.6%	39.6%	41.1%
	Engineering	13.0%	7.6%	10.3%
	Humanities	9.3%	11.3%	10.3%

5.1. Descriptive Statistics

Subsequently, customer rating distributions are denoted as follows: For the triple of customer rating distributions in each decision, AM_1 is the distribution $(x_1, \ldots, x_5)$ for which the arithmetic mean $am_1 = \sum_{k=1}^{5} k \cdot x_{k,1}$ is highest, AM_2 the one with the second-highest mean am_2, and AM_3 the one with the lowest mean value am_3. Thus, assuming that subjects rank products by the arithmetic mean, *[AM_1,AM_2,AM_3]* (i.e., $AM_1 \succ AM_2 \succ AM_3$) corresponds to the *reference ranking*.

Overall, subjects made $54 \times 12 = 648$ decisions in the information treatment and $53 \times 12 = 636$ decisions in the control treatment. Figure 6 shows the relative frequency distributions over the six possible rankings in both treatments. 44.3% of the decisions in the information treatment and 44.7% of the decisions in

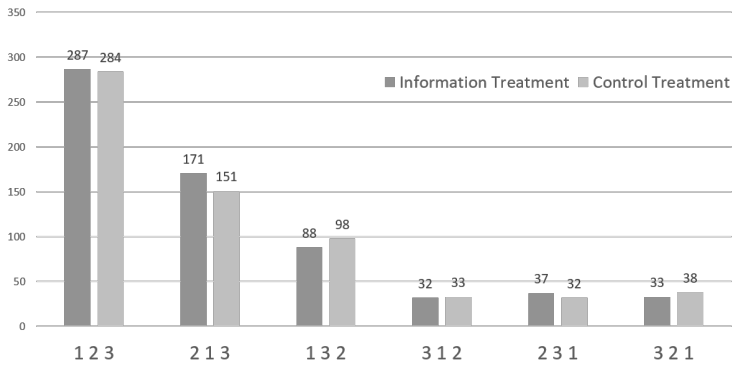


Figure 6: Absolute frequencies of observed rankings in the control and information treatment. *1* (abbreviation for AM_1) is the object with the highest mean, *2* the one with the second-highest mean, and *3* the one with the lowest mean. Hence, *1 2 3* corresponds to the ranking in accordance with the arithmetic mean.

the control treatment are in coherence with the arithmetic mean.[8] Measuring the distance between the reference ranking *[AM_1,AM_2,AM_3]* and any other ranking in terms of the Kendall distance, i.e., the number of pairwise inversions between objects (which means that the distance is between 0 and 3), 84.3% of all decisions in the information treatment and 83.8% in the control treatment are within a distance of at most

[8] Detailed ranking decisions are indicated over the 12 categories and both treatments in Table A.11.

1 from the reference ranking (i.e., *[AM$_1$,AM$_2$,AM$_3$]*, *[AM$_2$,AM$_1$,AM$_3$]*, and *[AM$_1$,AM$_3$,AM$_2$]*).

Considering also the aggregation functions described in Chapter 2.2.2, we compare the ranking behavior in Table 2 with regard to the different heuristics. We provide win/tie/loss statistics. Thereby, all pairs of heuristics are compared on each observed ranking by their Kendall distance (the one with the smaller distance "winning", the other one "losing"). As the arithmetic mean (AM) wins most often, these results constitute evidence in favor of the AM heuristic.

Table 2: Win/tie/loss statistics for the different heuristics, separately for control (top) and information treatment (bottom).

	AM	FIV	ONE	BIN	POS	NEG	MED
AM	0 / 636 / 0	308 / 286 / 42	191 / 330 / 115	243 / 231 / 162	200 / 372 / 64	184 / 317 / 135	237 / 371 / 38
FIV	42 / 286 / 308	0 / 636 / 0	207 / 28 / 401	157 / 134 / 345	144 / 194 / 298	181 / 64 / 391	150 / 270 / 216
ONE	115 / 330 / 191	401 / 28 / 207	0 / 636 / 0	229 / 204 / 203	263 / 198 / 175	121 / 403 / 112	326 / 127 / 183
BIN	162 / 231 / 243	345 / 134 / 157	203 / 204 / 229	0 / 636 / 0	265 / 212 / 159	155 / 265 / 216	227 / 309 / 100
POS	64 / 372 / 200	298 / 194 / 144	175 / 198 / 263	159 / 212 / 265	0 / 636 / 0	158 / 227 / 251	265 / 271 / 100
NEG	135 / 317 / 184	391 / 64 / 181	112 / 403 / 121	216 / 265 / 155	251 / 227 / 158	0 / 636 / 0	330 / 143 / 163
MED	28 / 371 / 237	216 / 270 / 150	183 / 127 / 326	100 / 309 / 227	100 / 271 / 265	163 / 143 / 330	0 / 636 / 0
AM	0 / 648 / 0	314 / 298 / 36	187 / 343 / 118	229 / 239 / 180	182 / 396 / 70	179 / 333 / 136	242 / 378 / 28
FIV	36 / 298 / 314	0 / 648 / 0	198 / 33 / 417	147 / 144 / 357	134 / 193 / 321	179 / 73 / 396	132 / 284 / 232
ONE	118 / 343 / 187	417 / 33 / 198	0 / 648 / 0	221 / 196 / 231	248 / 219 / 181	118 / 396 / 134	329 / 134 / 185
BIN	180 / 239 / 229	357 / 144 / 147	231 / 196 / 221	0 / 648 / 0	274 / 216 / 158	169 / 270 / 209	232 / 320 / 96
POS	70 / 396 / 182	321 / 193 / 134	181 / 219 / 248	158 / 216 / 274	0 / 648 / 0	169 / 235 / 244	270 / 278 / 100
NEG	136 / 333 / 179	396 / 73/ 179	134 / 396 / 118	209 / 270 / 169	244 / 235 / 169	0 / 648 / 0	332 / 150 / 166
MED	28 / 378 / 242	232 / 284 / 132	185 / 134 / 329	96 / 320 / 232	100 / 278 / 270	166 / 150 / 332	0 / 648 / 0

5.2. Overall Congruency of Heuristics with the Arithmetic Mean

We fitted our statistical model in a subject-wise setting. Every model is obtained using $|\mathcal{D}| = 12$ ranking instances consisting of three products each. The fitted models are then analyzed in order to find evidence in favor or against the propositions from Chapter 2. The preferences of the i^{th} subject can be characterized in terms of the corresponding parameter estimate $\boldsymbol{w}_i^*$.[9] In Table 3 the averaged preferences from both treatments are compared to the theoretical and estimated model parameters of the AM heuristic. We see that the observed behavior is very similar to the arithmetic mean aggregation function in both treatments.

Table 3: Average and the standard deviation of the estimated model parameters in both CT and IT groups and the corresponding values for the AM heuristic.

	w_1	w_2	w_3	w_4	w_5	α
CT	$-0.315 \pm .116$	$-0.104 \pm .148$	$0.002 \pm .086$	$0.169 \pm .085$	$0.248 \pm .14$	89.041 ± 90.2
IT	$-0.317 \pm .109$	$-0.094 \pm .166$	$0.02 \pm .088$	$0.167 \pm .106$	$0.223 \pm .139$	105.071 ± 103.9
AM	-0.333	-0.167	0.000	0.167	0.333	
Estimated AM	-0.377	-0.123	0.024	0.155	0.321	

[9] The estimated preference vector together with the precision parameter α_i are provided in Tables B.13 and B.14. Figure B.10 shows a graphical representation of these parameters, also including other heuristics.

We measure the Kendall distance between a product ranking predicted by a fitted model and the ground truth ranking given by participants' heuristics. The averaged distances are provided in Table 4. As expected, the estimated models perform best due to the higher model flexibility. However, the performance of the AM heuristic is better than any other heuristic and is only slightly outperformed by the fitted model. This finding appears to be consistent with Proposition 1.

Table 4: Average Kendall distance for the different heuristics and the estimated model (separately for control and information treatment).

	AM	FIV	ONE	BIN	POS	NEG	MED	estimated
CT	0.775	1.509	0.981	1.038	1.204	0.942	1.437	0.596
IT	0.765	1.515	0.981	0.957	1.136	0.895	1.380	0.568

Goodness-of-fit. We use all data for model fitting, making it difficult to evaluate goodness-of-fit directly due to absence of test data. However, we can indirectly evaluate the goodness-of-fit of our model by comparing its results with results from other methodological approaches: In comparison to our win/tie/loss analysis, we find strong evidence for a good fit of our model (cf. Table B.17). In addition to the preceding model, we compute the likelihood ratio statistic Λ between various reference heuristics (being the null hypothesis) and the fitted model (maximum likelihood estimate). The statistics and the corresponding p-value for every reference aggregation function are provided in Tables B.15 and B.16. For informativeness, we average the statistic values across the heuristic in Table B.18. Although the p-values are not significant in many cases, the lower Λ statistic values of the AM heuristic indicate consistent results of the win/tie/loss statistics and the parametric estimates.

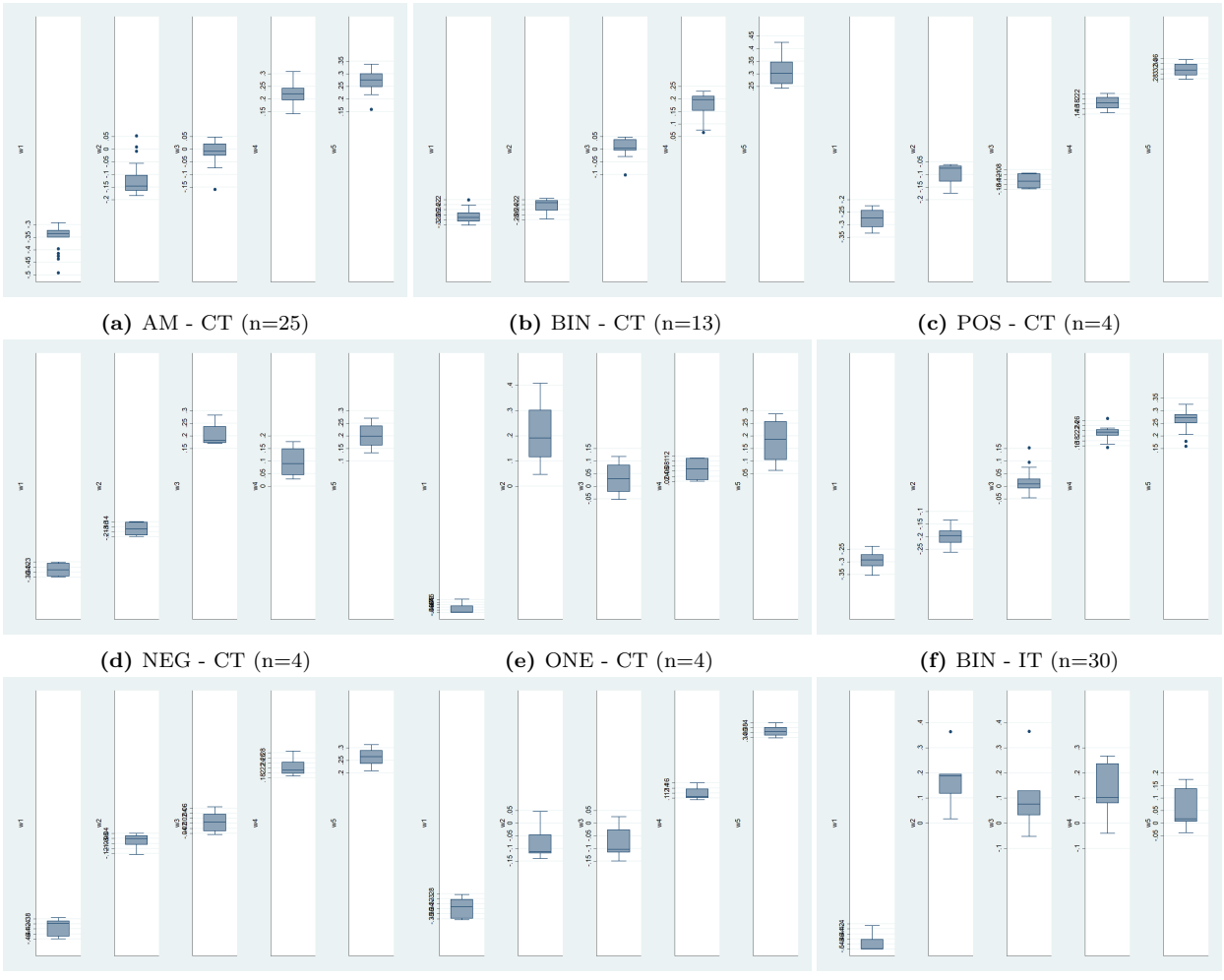
Concluding the preceding analysis, we find support for Proposition 1:

Result 1. *Provided with product rating distributions to deduce the quality of products, the average behavior of the subjects is described best by the arithmetic mean.*

5.3. Heterogeneity in Aggregation Heuristics

So far, our analysis does not consider heterogeneity in the data that might be explained by systematic patterns. Thereby, the parameter estimate $\boldsymbol{w}_i^*$ allows us to define a dissimilarity $d(i,j) = \|\boldsymbol{w}_i^* - \boldsymbol{w}_j^*\|$ between subjects that we can use to cluster subjects, using agglomerative hierarchical weighted-average clustering. Details can be found in the Appendix. Processing the information from the identified clusters (cf. Table C.19, C.20, and C.21), we find five major clusters in the control treatment and four major clusters in information treatment. Clusters are visualized in Figure 7.

In the control treatment, the largest cluster contains subjects whose decisions are consistent with the arithmetic mean, at least approximately (AM, n=25). The second major cluster reveals the tendency to



(a) AM - CT (n=25) **(b)** BIN - CT (n=13) **(c)** POS - CT (n=4)

(d) NEG - CT (n=4) **(e)** ONE - CT (n=4) **(f)** BIN - IT (n=30)

(g) AM - IT (n=7) **(h)** AM/POS - IT (n=5) **(i)** ONE - IT(n=5)

Figure 7: Boxplots of category weights w_1 to w_5 for various clusters in control treatment (cf. Figures (a) - (e)) and information treatment (cf. Figures (f) - (i)).

Table 5: Deviations from arithmetic mean reference category weights. + indicates positive deviations, − negative deviations. Asterisks show significance levels of deviations.

		w1	w2	w3	w4	w5
AM*		-0.377	-0.123	0.024	0.155	0.321
AM - CT	25	$+^{**}$	0	$-^{***}$	$+^{***}$	$-^{***}$
BIN - CT	13	$+^{***}$	$-^{***}$	0	0	0
BIN - IT	30	$+^{***}$	$-^{***}$	0	$+^{***}$	$-^{***}$
AM - IT	7	$-^{**}$	$+^{**}$	0	$+^{**}$	$-^{**}$

Table 6: Deviations from binary reference category weights. + indicates positive deviations, − negative deviations. Asterisks show significance levels of deviations.

		w1	w2	w3	w4	w5
BIN*		-0.236	-0.264	0.021	0.232	0.247
AM - CT	25	$-^{***}$	$+^{***}$	$-^{***}$	$-^{**}$	$+^{***}$
BIN - CT	13	$-^{***}$	$+^{***}$	0	$-^{***}$	$+^{***}$
BIN - IT	30	$-^{***}$	$+^{***}$	0	$-^{***}$	$+^{***}$
AM - IT	7	$-^{**}$	$+^{**}$	0	0	0

estimate the positive categories w_4 and w_5, respectively negative categories w_1 and w_2, binarily (BIN, n=13). Smaller clusters focus on minimizing negative ratings (NEG, n=4), minimizing 1-star ratings (ONE, n=4), or maximizing positive ratings (POS, n=4).

In the information treatment, in which the arithmetic mean values and the relative frequencies are provided, we find for the majority the tendency to overweight moderate categories w_2 and w_4 and underweight extreme categories w_1 and w_5 (BIN, n=30). Another cluster chooses rather in accordance with the arithmetic mean (AM, n=7). The third cluster weights w_2 and w_3 similarly, resulting in a mixture of POS and AM (POS-AM; n=5). Also, in the information treatment, a small cluster focuses on the 1-star category (ONE, n=5).

Given the magnitude of BIN, we investigate whether there are systematic deviations from empirically estimated AM and BIN weights (cf. Table A.12). Results of the Wilcoxon signed-rank test (cf. Tables 5 and 6) suggest category weighting that is rather a mixture of the AM and BIN heuristic for major clusters that contain reference heuristics AM* and BIN* in the control treatment and information treatment. Figure 8 provides the median category weights of these clusters and illustrates that most of the category weights are indeed mixtures of AM* and BIN*.

Overall, we see the pattern of weighting in accordance with ordinal categories. However, the behavior show deviations from the arithmetic mean with its cardinal assumptions. The binary pattern is prevalent in both treatments. Additionally, small clusters focus only on 1-star ratings in both treatments.

Result 2. *There is heterogeneity in aggregation behavior. The majority aggregates customer rating distributions by employing the arithmetic mean with binary biases. In addition, minorities focus on negative or*

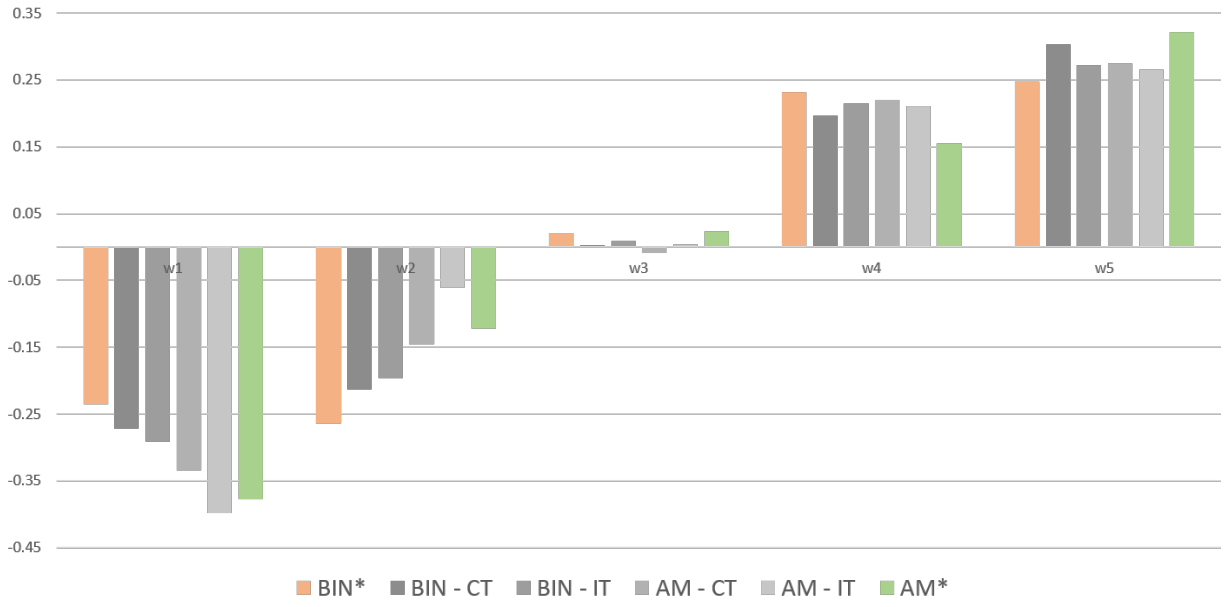


Figure 8: Median category weights w_1, ..., w_5 separated for AM and BIN clusters of CT and IT. Major clusters show evidence for mixtures of both heuristics.

worst rating categories.

5.4. Impact of Numerical Information

In both treatments, subjects are provided with graphical information of the customer ratings. In the information treatment, additional numerical information is provided. We test Proposition 3, which states that the numerical information affects behavior and, hence, cateogory weights. Comparing category weights between treatments with the Mann-Whitney-U-test, we do not find significant differences ($p = 0.25$ or larger) for any category weights.[10] Figure 9 depicts the similarities between treatments.

Result 3. *Overall, the additional provision of numerical information does not affect subjects' applied heuristics to deduce the quality of products.*

5.5. Determinants of Employed Heuristics

We follow two strings of analyses in this chapter. First, we investigate whether employing the arithmetic mean is correlated with individual characteristics. Secondly, we more generally examine whether the identified clusters contain subjects with specific individual characteristics.

[10] w_1 : $z = -.346$, $p = .7295$, w_2 : $z = .212$, $p = .8322$, w_3 : $z = -.897$, $p = .3696$, w_4 : $z = -.636$, $p = .5251$, w_5 : $z = 1.128$, $p = .2594$)

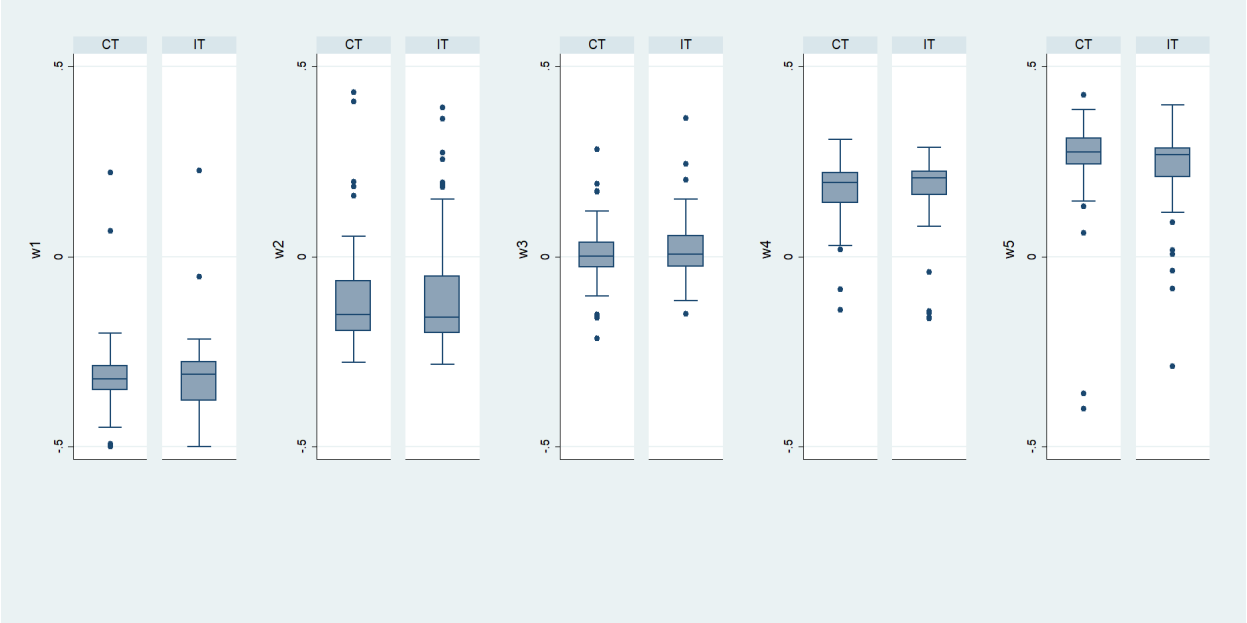


Figure 9: Boxplots of category weights w_1, ..., w_5 separated by treatments. There is no evidence for differences driven by provision of numerical information.

The impact of individual characteristics. We test for correlations between individual characteristics: i.e., age, gender, and studies with the estimates of the α^{Mean}-values. We pool data from both treatments, as we do not find significant differences between both treatments with regard to the α^{Mean}-values (MWU: $p = 1.000,\ z = 0.000$). We separate participants into two groups and compare the congruency to the arithmetic mean, i.e., the α^{Mean}-values, between those groups with the Mann-Whitney-U test. The field of studies is compared pairwise between the four groups of economics, education, engineering, and humanities. We use the median values to separate older from younger participants (median age: 22Y), respectively first-year students from more experienced students (median: 3 semesters). Female subjects are weakly significantly more conform with the arithmetic mean (MWU: $p = 0.055,\ z = -1.919$). Age, process in, and the field of studies do not have a significant effect on conformity with the arithmetic mean (cf. Table D.22 for test statistics). We identified four clusters in the information treatment and five clusters in the control treatment. We also conduct analyses across these clusters, but do not find significant effects of individual characteristics on the clusters (cf. Table D.23 in the Appendix for test statistics).

The impact of experience. Considering the influence of experience in online shopping on the employed heuristics, we separate participants in two groups. *Experienced subjects* (n=28) are defined as frequent buyers, i.e., weekly or more, who state the relevance of online reviews as important or rather important. Residuals are classified as *non-experienced* (n=79).[11] First, we investigate the congruency with the arith-

[11] This measure of experience is independent of the treatment ($\chi^2 : z = 0.2475, p = 0.619$).

metic mean by comparing α^{Mean}-values between both groups. On average, there is the tendency, though no statistical significance, that experienced subjects show a higher congruency with the arithmetic mean ($\alpha^{Mean}_{Experienced} = 41.80$ (22.72), $\alpha^{Mean}_{Non-Experienced} = 34.32$ (19.41), MWU: $z = -1.566,\ p = 0.1173$).[12] Second, we analyze whether there is a difference between clusters with regard to experience in online shopping. However, we do not find significant differences between clusters with regard to experience (cf. Table D.24 for test statistics).

The impact of risk attitudes. Analyzing the influence of risk attitudes on the employed heuristics, we use three measures of risk: risk-seeking with regard to trusting other people, risk-seeking in economic decisions, and the general risk attitude. Therefore, we calculate median values of these variables and classify subjects as risk-seeking when their statements are above the median risk value. First, we focus on the behavior in accordance with the arithmetic mean and do not find an effect on decision making for any of the three measures of risk attitudes.[13] Also, the analysis of the clusters with regard to different risk attitudes provides no evidence for an impact of risk attitudes on aggregation behavior (cf. Table D.25 for test statistics).

Hence, we conclude:

Result 4. *Women are more likely to choose in accordance with the arithmetic mean. Overall, however, decision making is not explained by individual characteristics, experience in online shopping or risk attitudes.*

5.6. Further Results

Congruence of self-claimed and observed aggregation behavior. After the experiment we asked subjects about their decision criteria for assessing customer rating distributions. Multiple answers were allowed. In the questionnaire, participants state the aggregation heuristics depicted in Table 7. We find treatment effects

Table 7: Decision criteria used by the participants according to the answers in the questionnaire.

Criteria	Information Treatment	Control Treatment
Arithmetic Mean	59.3%	39.6%
Median	13.0%	13.2%
Five Stars	68.5%	54.7%
One Star	61.1%	56.6%
Negative Ratings	50.0%	49.1%
Positive Ratings	55.6%	67.9%
No Criteria	0%	7.5%

for the answers regarding the arithmetic mean ($\chi^2:\ z = 4.1259,\ p = 0.042$) and no criteria ($\chi^2:\ z = 4.2337,\ p = 0.040$). When subjects have more information about options (in the IT) they behave more in line with the self-claimed strategies that were elicited in questionnaires after the experiment (cf. Table 8).

[12] Repeating analysis with Kendall's distance as the dependent variable we find weakly significant differences.

[13] $Risk_{General}:\ z = 0.449,\ p = 0.6536$, $Risk_{people}:\ z = -.0584,\ p = 0.5590$, $Risk_{Economic}:\ z = -0.489,\ p = 0.6251$.

Without the numerical information, the self-claimed strategies do not correspond to the observed behavior (in the CT, cf. Table 9).

Result 5. *On average, self-claimed aggregation heuristics correspond to behavior when numerical information is provided. In the absence of numerical information, self-claimed answers on aggregation heuristics do not match behavior.*

Table 8: Information Treatment - no attitude behavior gap.

Strategy	Claimed	Mean deviation	Std.dev.	n	MWU: z-value	p-value
AM-Strategy	yes	8.188	4.468	32	1.884	0.0596
	no	10.636	5.499	22		
MED-Strategy	yes	12.857	2.116	7	0.523	0.6010
	no	13.319	1.889	47		
POS-Strategy	yes	12.633	3.178	30	2.055	0.0398
	no	14.875	4.100	24		
NEG-Strategy	yes	10.111	4.089	27	1.488	0.1369
	no	11.370	3.176	27		
ONE-Strategy	yes	11.485	2.412	33	0.286	0.7747
	no	12.238	3.477	21		
FIV-Strategy	yes	17.378	3.759	37	3.470	0.0005
	no	21.235	2.796	41		

Table 9: Control Treatment - attitude behavior gap.

Strategy	Claimed	Mean deviation	Std.dev.	n	MWU: z-value	p-value
AM-Strategy	yes	8.714	4.971	21	0.950	0.3419
	no	9.688	4.314	32		
MED-Strategy	yes	15.429	1.512	7	-2.105	0.0353
	no	13.239	2.758	46		
POS-Strategy	yes	14.111	3.196	36	1.346	0.1783
	no	15.176	2.157	17		
NEG-Strategy	yes	10.692	3.082	26	1.326	0.1848
	no	11.889	3.856	27		
ONE-Strategy	yes	11.200	3.537	30	1.183	0.2370
	no	12.522	3.788	23		
FIV-Strategy	yes	18.621	3.458	29	-0.396	0.6921
	no	18.000	4.364	24		

6. Conclusion

Aggregation mechanisms are implemented in reputation systems to remedy information overload. However it is unclear whether the technically implemented aggregation functions are in accordance with the actual aggregation behavior of customers. Hence, we conducted a laboratory experiment to elicit subjects' aggregation patterns and compare these to reference aggregation functions.

Overall, we find evidence that the arithmetic mean is an appropriate aggregation function. However, our analysis of the major clusters reveals the tendency to overweight moderate ratings and underweight extreme ratings, thus indicating the binary bias (cf. Fisher et al. 2018). Additionally, minor clusters focus on customer rating distributions with the least 1-star ratings or the least negative (i.e., 1- and 2-star) ratings. Overall though, contrary to predictions, the individual characteristics, risk attitudes, or experience in online shopping do not affect the employed aggregation heuristics as we only identify that women decide weakly significant more in accordance with the arithmetic mean. The employed heuristics are also not affected systematically by the treatment variations (only visualization or visualization enriched by numerical information). This also indicates that the aforementioned binary bias is rather a conscious pattern and not a behavioral bias driven by bounded cognitive abilities.

Our research is limited with regard to the considered classes of aggregation functions. In particular, we only considered simultaneous criteria heuristics. Hauser (2014, p.1692) argue that heuristic decision rules can also be sequential. We also abstract from the dimension of time, which is also important in aggregation metrics. For example, Ivanova and Scholz (2017) propose a dynamic approach, aggregating the recent ratings to k-values and thereby reducing the influence of fake reviews. Considering the time trends in customer ratings and possible quality shifts, Leberknight et al. (2011) propose employing the Average Rating Volatility (ARV) to derive a better aggregation measure. Dai et al. (2018) use data from Yelp.com to generate a better aggregation function, allowing for misjudgments in quality, information cascades regarding customer reviews, and quality variation over time. Their result shows the advantages of a simple algorithm compared to the arithmetic mean.

Nevertheless, this article has important implications. Independent of whether numerical information is provided or not, we identify a systematic binary bias in aggregation behavior that is not sufficiently considered in practice. Given the j-shape of customer rating distributions, the binary bias of customers can lead to severe deviations between the calculated valence (by reputation systems) and the perceived valence of customers that might result in inefficiencies. We also identify heterogeneity in aggregation behavior. Minor clusters focused on the minimizing of 1-star ratings, negative ratings, or positive ratings. Addressing this heterogeneity, one could enhance reputation systems by calculating personalized valence values. Serving as a role model to elicit suitable aggregation functions and personalize provided valence values, this can be achieved by implementing our design in online marketplaces.

Our methodological approach of eliciting heuristics also highlights the sensitivity of empirical studies employing questionnaires. Although we do not find differences in ranking behavior, answers in the questionnaire differ significantly between both treatments. Thereby, the analysis of our questionnaire additionally shows that customers can phrase their strategies better when they are provided with additional numerical information. Our study also provides evidence for the appropriateness of data-driven approaches. By employing the Plackett-Luce model with only basic assumptions, we however achieve results that are plausible

and verifiable.

Acknowledgements

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre 901 "On-The-Fly Computing" (SFB 901) under the project number 160364472-SFB901.

References

Adomavicius, G., Bockstedt, J. C., Curley, S. P., and Zhang, J. (2019). Reducing recommender system biases: An investigation of rating display designs. *MIS Quarterly: Management Information Systems*, 43(4):1321–1341.

Askalidis, G., Kim, S. J., and Malthouse, E. C. (2017). Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:23–30.

Bae, J. and Koo, D.-M. (2018). Lemons problem in collaborative consumption platforms: Different decision heuristics chosen by consumers with different cognitive styles. *Internet Research*, 28(3):746–766.

Beliakov, G., Calvo, T., and James, S. (2011). Aggregation of preferences in recommender systems. In *Recommender systems handbook*, pages 705–734. Springer.

Besedeš, T., Deck, C., Sarangi, S., and Shor, M. (2012). Age effects and heuristics in decision making. *Review of Economics and Statistics*, 94(2):580–595.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208.

Camilleri, A. R. (2017). The presentation format of review score information influences consumer preferences through the attribution of outlier reviews. *Journal of Interactive Marketing*, 39:1–14.

Chen, P.-Y., Hong, Y., and Liu, Y. (2018). The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science*.

Cheng, W., Hüllermeier, E., and Dembczynski, K. J. (2010). Label ranking methods based on the plackett-luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222.

Cheung, C. M. and Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision support systems*, 53(1):218–225.

Cheung, C. M. and Thadani, D. R. (2012). The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision support systems*, 54(1):461–470.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Coba, L., Rook, L., Zanker, M., and Symeonidis, P. (2019). Decision making strategies differ in the presence of collaborative explanations: Two conjoint studies. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 291–302. ACM.

Dai, W., Jin, G., Lee, J., and Luca, M. (2018). Aggregation of consumer ratings: an application to yelp.com. *Quantitative Marketing and Economics*, 16(3):289–339.

De Langhe, B., Fernbach, P. M., and Lichtenstein, D. R. (2015). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424.

Dellarocas, C. (2005). Reputation mechanism design in online trading environments with pure moral hazard. *Information systems research*, 16(2):209–230.

Duan, W., Gu, B., and Whinston, A. B. (2008). Do online reviews matter?–an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fisher, M., Newman, G. E., and Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, 45(3):471–489.

Flanagin, A. J., Metzger, M. J., Pure, R., and Markov, A. (2011). User-generated ratings and the evaluation of credibility and product quality in ecommerce transactions. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., and Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, 90(2):217–232.

Garcin, F., Faltings, B., and Jurca, R., editors (2009). *Aggregating reputation feedback*, volume 1.

Garcin, F., Xia, L., and Faltings, B. (2013). How aggregators influence human rater behavior. In *Proc. Workshop@ 14th ACM Conference on Electronic Commerce (EC-13)*.

Gigerenzer, G. and Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart*, pages 3–34. Oxford University Press.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., and Chen, J. (2019). Design of review systems–a strategic instrument to shape online reviewing behavior and economic outcomes. *The Journal of Strategic Information Systems*.

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, 67(8):1688–1699.

He, S. X. and Bond, S. D. (2015). Why is the crowd divided? attribution for dispersion in online word of mouth. *Journal of Consumer Research*, 41(6):1509–1527.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing*, 18(1):38–52.

Ho, Y.-C., Wu, J., and Tan, Y. (2017). Disconfirmation effect on online rating behavior: A structural model. *Information Systems Research*, 28(3):626–642.

Hong, H., Xu, D., Wang, G. A., and Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems*, 102:1–11.

Hoyer, B. and van Straaten, D. (2021). Anonymity and Self-Expression in Online Rating Systems - An Experimental Analysis. Working Papers Dissertations 70, Paderborn University, Faculty of Business Administration and Economics.

Hu, N., Pavlou, P. A., and Zhang, J. (2006). Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 324–330. ACM.

Hu, N., Pavlou, P. A., and Zhang, J. J. (2017). On self-selection biases in online product reviews. *MIS Quarterly*, 41(2):449–471.

Hu, N., Zhang, J., and Pavlou, P. A. (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM*, 52(10):144–147.

Hurley, R. F. and Estelami, H. (1998). Alternative indexes for monitoring customer perceptions of service quality: A comparative evaluation in a retail context. *Journal of the academy of Marketing Science*, 26(3):209–221.

Ivanova, O. and Scholz, M. (2017). How can online marketplaces reduce rating manipulation? a new approach on dynamic aggregation of online ratings. *Decision Support Systems*, 104:64–78.

Josang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644.

Jurca, R., Garcin, F., Talwar, A., and Faltings, B. (2010). Reporting incentives and biases in online review forums. *ACM Transactions on the Web (TWEB)*, 4(2):5.

Kostyk, A., Niculescu, M., and Leonhardt, J. M. (2017). Less is more: Online consumer ratings' format affects purchase intentions and processing. *Journal of Consumer Behaviour*, 16(5):434–441.

Leberknight, C. S., Sen, S., and Chiang, M. (2011). On the volatility of online ratings: an empirical study. In *Workshop on E-Business*, pages 77–86. Springer.

Lee, M., Rodgers, S., and Kim, M. (2009). Effects of valence and extremity of ewom on attitude toward the brand and website. *Journal of Current Issues & Research in Advertising*, 31(2):1–11.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA.

Marinescu, I., Klein, N., Chamberlain, A., and Smart, M. (2018). Incentives can reduce bias in online reviews. Technical report, National Bureau of Economic Research.

McGlohon, M., Glance, N., and Reiter, Z. (2010). Star quality: Aggregating reviews to rank products and merchants. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Metzger, M. J. and Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59:210–220.

Metzger, M. J., Flanagin, A. J., and Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of communication*, 60(3):413–439.

Mousavi, S. and Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. *Journal of Business Research*, 67(8):1671 – 1678.

Munzel, A. and H. Kunz, W. (2014). Creators, multipliers, and lurkers: who contributes and who benefits at online review sites. *Journal of Service Management*, 25(1):49–74.

Peddibhotla, N. B. and Subramani, M. R. (2007). Contributing to public document repositories: A critical mass theory perspective. *Organization Studies*, 28(3):327–346.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202.

Qiu, L., Pang, J., and Lim, K. H. (2012). Effects of conflicting aggregated rating on ewom review credibility and diagnosticity: The moderating role of review valence. *Decision Support Systems*, 54(1):631–643.

Reyes, R. M., Thompson, W. C., and Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology*, 39(1):2.

Rozenkrants, B., Wheeler, S. C., and Shiv, B. (2017). Self-expression cues in product rating distributions: When people prefer polarizing products. *Journal of Consumer Research*, 44(4):759–777.

Schäfer, D. and Hüllermeier, E. (2018). Dyad ranking using plackett–luce models based on joint feature representations. *Machine Learning*, 107(5):903–941.

Schoenmueller, V., Netzer, O., and Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research*, 57(5):853–877.

Sharot, T. and Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in cognitive sciences*, 20(1):25–33.

Sundar, S. S., Oeldorf-Hirsch, A., and Xu, Q. (2008). The bandwagon effect of collaborative filtering technology. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3453–3458, New York, NY, USA. ACM.

Sundar, S. S., Xu, Q., and Oeldorf-Hirsch, A. (2009). Authority vs. peer: how interface cues influence users. In *CHI'09 Extended Abstracts on human factors in computing systems*, pages 4231–4236. ACM.

Tam, K. Y. and Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information systems research*, 16(3):271–291.

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

van Straaten, D. (2021). Incentive Schemes in Customer Rating Systems - Comparing the Effects of Unconditional and Conditional Rebates on Intrinsic Motivation. Working Papers Dissertations 71, Paderborn University, Faculty of Business Administration and Economics.

von Helversen, B., Abramczuk, K., Kopeć, W., and Nielek, R. (2018). Influence of consumer reviews on online purchasing decisions in older and younger adults. *Decision Support Systems*, 113:1–10.

Watson, J., Ghosh, A. P., and Trusov, M. (2018). Swayed by the numbers: the consequences of displaying product review attributes. *Journal of Marketing*, 82(6):109–131.

Wu, P. F. (2019). Motivation crowding in online product reviewing: A qualitative study of amazon reviewers. *Information & Management*.

Wu, W.-Y. and Chang, M.-L. (2007). The role of risk attitude on online shopping: Experience, customer satisfaction, and repurchase intention. *Social Behavior and Personality: an international journal*, 35(4):453–468.

Wulff, D. U., Hills, T. T., and Hertwig, R. (2015). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*, 28(3):214–223.

Yang, X., Yang, G., and Wu, J. (2016). Integrating rich and heterogeneous information to design a ranking system for multiple products. *Decision Support Systems*, 84:117–133.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69(3):237–249.

Yin, C., Sun, Y., Fang, Y., and Lim, K. (2018). Exploring the dual-role of cognitive heuristics and the moderating effect of gender in microblog information credibility evaluation. *Information Technology & People*.

You, Y., Vadakkepatt, G. G., and Joshi, A. M. (2015). A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, 79(2):19–39.

Zhang, K. Z., Zhao, S. J., Cheung, C. M., and Lee, M. K. (2014). Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems*, 67:78–89.

Zimmermann, S., Herrmann, P., Kundisch, D., and Nault, B. R. (2018). Decomposing the variance of consumer ratings and the impact on price and demand. *Information Systems Research*, 29(4):984–1002.

Appendix A.

Table A.10: Product ratings distributions combined into the product groups.

product	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	Mean value	Median
0	0.086	0.028	0.044	0.126	0.716	4.358	5
1	0.074	0.025	0.074	0.222	0.605	4.259	5
2	0.069	0.080	0.103	0.149	0.599	4.129	5
3	0.038	0.055	0.066	0.159	0.682	4.392	5
4	0.027	0.054	0.081	0.257	0.581	4.311	5
5	0.102	0.082	0.020	0.061	0.735	4.245	5
6	0.154	0.077	0.115	0.038	0.616	3.885	5
7	0.149	0.085	0.064	0.213	0.489	3.808	4
8	0.162	0.030	0.131	0.273	0.404	3.727	4
9	0.140	0.065	0.061	0.170	0.564	3.953	5
10	0.021	0.191	0.120	0.291	0.377	3.812	4
11	0.061	0.087	0.122	0.700	0.030	3.551	4
12	0.017	0.089	0.400	0.106	0.388	3.759	3
13	0.056	0.133	0.193	0.346	0.272	3.645	4
14	0.240	0.111	0.135	0.005	0.509	3.432	5
15	0.000	0.048	0.504	0.087	0.361	3.761	3
16	0.195	0.012	0.145	0.259	0.389	3.635	4
17	0.246	0.150	0.091	0.012	0.501	3.372	5
18	0.017	0.276	0.208	0.020	0.479	3.668	3
19	0.175	0.058	0.149	0.293	0.325	3.535	4
20	0.335	0.012	0.134	0.007	0.512	3.349	5
21	0.034	0.211	0.256	0.076	0.423	3.643	3
22	0.249	0.052	0.017	0.367	0.315	3.447	4
23	0.411	0.012	0.074	0.000	0.503	3.172	5
24	0.290	0.216	0.000	0.129	0.365	3.063	2
25	0.262	0.030	0.402	0.164	0.142	2.894	3
26	0.447	0.000	0.048	0.384	0.121	2.732	4
27	0.165	0.074	0.057	0.102	0.602	3.902	5
28	0.255	0.053	0.051	0.119	0.522	3.600	5
29	0.214	0.071	0.143	0.143	0.429	3.502	4
30	0.059	0.029	0.029	0.147	0.736	4.472	5
31	0.097	0.065	0.000	0.129	0.709	4.288	5
32	0.143	0.024	0.000	0.119	0.714	4.237	5
33	0.039	0.039	0.066	0.158	0.698	4.437	5
34	0.097	0.065	0.065	0.065	0.708	4.222	5
35	0.115	0.082	0.033	0.180	0.590	4.048	5

Table A.11: Ranking decisions in information and control treatments. R1 is the ranking with customer rating distributions being ranked in accordance with the arithmetic mean (1st, 2nd, 3rd). R2 is the ranking (2nd, 1st, 3rd). R3 is the ranking (1st, 3rd, 2nd). R4 is the ranking (3rd, 1st, 2nd). R5 is the ranking (2nd, 3rd, 1st). R6 is the ranking (3rd, 2nd, 1st). See Table A.10 for a list of options.

Decision	Treatment	R1	R2	R3	R4	R5	R6
1 (Options 1-3)	IT	23	2	21	2	0	6
	CT	19	1	25	4	0	4
2 (Options 4-6)	IT	45	0	8	0	0	1
	CT	43	1	6	1	2	0
3 (Options 7-9)	IT	39	3	3	5	1	3
	CT	27	4	13	5	2	2
4 (Options 10-12)	IT	21	13	11	4	3	2
	CT	21	18	4	5	1	4
5 (Options 13-15)	IT	31	9	3	2	3	6
	CT	22	13	6	3	4	5
6 (Options 16-18)	IT	24	7	5	1	10	7
	CT	23	4	9	2	5	10
7 (Options 19-21)	IT	8	8	20	11	3	4
	CT	16	6	15	6	4	6
8 (Options 22-24)	IT	13	32	2	3	3	1
	CT	17	27	4	1	2	2
9 (Options 25-27)	IT	22	17	7	2	6	0
	CT	24	15	3	2	6	3
10 (Options 28-30)	IT	11	37	1	0	3	2
	CT	18	28	2	1	2	2
11 (Options 31-33)	IT	29	20	5	0	0	0
	CT	29	17	3	1	3	0
12 (Options 34-36)	IT	21	23	2	2	5	1
	CT	25	17	8	2	1	0

Table A.12: Category weights of reference functions. Estimated category weights are provided in *-rows.

Heuristic		w1	w2	w3	w4	w5
AM		-0.333	-0.167	0.000	0.167	0.333
	*	-0.377	-0.123	0.024	0.155	0.321
FIV		-0.125	-0.125	-0.125	-0.125	0.500
	*	-0.056	-0.239	-0.072	-0.133	0.500
ONE		-0.500	0.125	0.125	0.125	0.125
	*	-0.500	0.259	0.139	0.068	0.034
BIN		-0.250	-0.250	0.000	0.250	0.250
	*	-0.236	-0.264	0.021	0.232	0.247
POS		-0.167	-0.167	-0.166	0.250	0.250
	*	-0.166	-0.168	-0.166	0.245	0.255
NEG		-0.250	-0.250	0.166	0.167	0.167
	*	-0.259	-0.241	0.161	0.161	0.178
MED		n.a.	n.a.	n.a.	n.a.	n.a.
	*	-0.006	-0.448	-0.046	0.122	0.378

Appendix B.

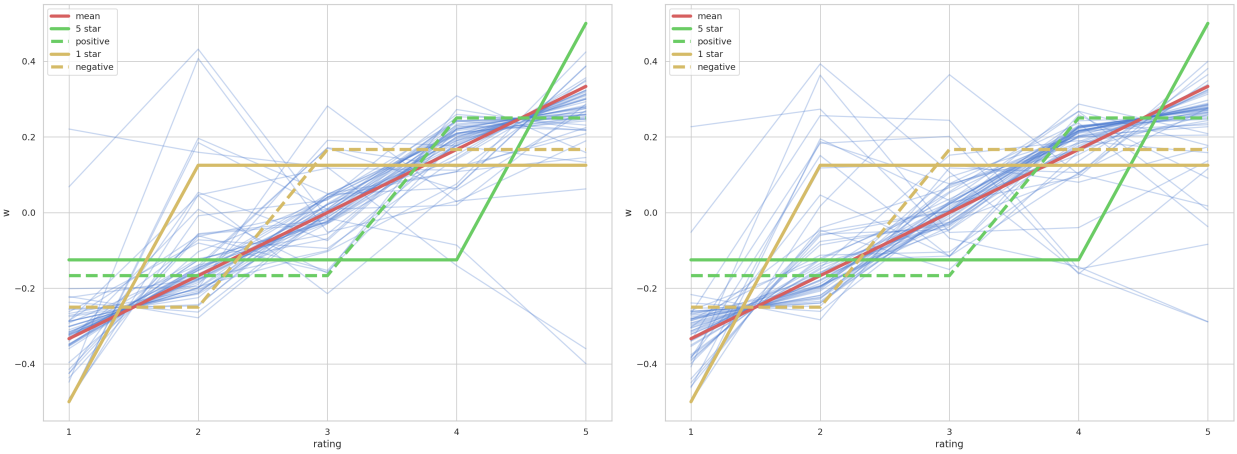


Figure B.10: Graphical representation of the parameter estimates $\boldsymbol{w}_i^*$ in the subject-wise setting (CT left, IT right). The diagonal line in red corresponds to the arithmetic mean reference weights.

Table B.13: Estimated model parameters for the participants in the control treatment group.

ID	w1	w2	w3	w4	w5	alpha
1	-0.286	0.045	-0.214	0.067	0.387	37.351
2	-0.237	-0.263	0.002	0.185	0.314	69.093
3	-0.201	-0.196	-0.103	0.076	0.424	38.519
4	-0.225	-0.175	-0.100	0.181	0.319	19.524
5	-0.349	-0.126	-0.025	0.253	0.247	88.157
6	-0.262	-0.080	-0.157	0.189	0.311	11.564
7	-0.448	0.407	-0.052	0.030	0.063	11.670
8	-0.323	-0.177	0.021	0.166	0.313	62.742
9	-0.315	-0.185	0.010	0.259	0.231	29.644
10	-0.360	-0.140	0.171	0.060	0.269	317.910
11	-0.321	-0.122	-0.057	0.220	0.280	236.074
12	-0.492	-0.008	0.033	0.308	0.159	62.814
13	-0.320	-0.180	0.000	0.243	0.257	62.197
14	-0.426	0.053	-0.074	0.229	0.218	75.952
15	-0.272	-0.216	-0.013	0.207	0.293	328.677
16	-0.286	-0.061	-0.153	0.222	0.278	37.775
17	-0.322	-0.178	0.006	0.195	0.299	54.877
18	-0.415	-0.058	-0.027	0.200	0.300	61.966
19	-0.222	-0.278	0.000	0.178	0.322	52.272
20	-0.351	-0.149	0.048	0.205	0.247	333.948
21	-0.288	-0.208	-0.005	0.197	0.303	58.133
22	-0.500	0.185	0.009	0.019	0.287	6.683
23	-0.351	-0.126	-0.023	0.191	0.309	47.823
24	-0.348	-0.152	0.007	0.209	0.284	56.325
25	-0.396	-0.104	0.035	0.190	0.275	83.110
26	-0.335	-0.165	0.040	0.209	0.251	341.454
27	-0.340	0.009	-0.160	0.272	0.219	48.326
28	-0.318	-0.153	-0.028	0.177	0.323	55.953
29	-0.300	-0.200	0.282	0.029	0.190	41.542
30	-0.310	-0.190	0.191	0.176	0.133	18.088
31	0.221	0.159	0.120	-0.140	-0.360	31.226
32	-0.256	-0.244	0.024	0.216	0.260	45.269
33	-0.292	-0.139	-0.069	0.234	0.266	147.960
34	-0.500	0.196	0.050	0.108	0.145	12.473
35	-0.301	-0.199	0.003	0.230	0.266	119.796
36	-0.267	-0.204	-0.029	0.153	0.347	100.813
37	-0.302	-0.198	0.038	0.208	0.254	183.887
38	-0.275	-0.225	0.047	0.210	0.242	52.949
39	-0.315	-0.176	-0.009	0.240	0.260	59.669
40	-0.437	-0.063	0.010	0.243	0.247	172.801
41	-0.425	-0.056	-0.019	0.200	0.300	30.864
42	-0.347	-0.153	0.041	0.242	0.217	339.314
43	0.068	0.432	-0.015	-0.086	-0.399	40.960
44	-0.288	-0.212	0.021	0.220	0.259	49.138
45	-0.325	-0.162	-0.014	0.220	0.280	148.532
46	-0.353	-0.147	0.171	0.120	0.208	57.246
47	-0.331	-0.145	-0.024	0.176	0.324	18.877
48	-0.333	-0.072	-0.095	0.144	0.356	90.039
49	-0.331	-0.146	-0.023	0.225	0.275	38.130
50	-0.327	-0.173	0.022	0.140	0.338	64.938
51	-0.287	-0.213	0.048	0.066	0.386	46.190
52	-0.253	-0.247	0.041	0.108	0.351	81.957
53	-0.500	0.046	0.117	0.112	0.224	35.983

Table B.14: Estimated model parameters for the participants in the information treatment group.

ID	w1	w2	w3	w4	w5	alpha
54	-0.286	-0.168	-0.047	0.183	0.317	105.542
55	-0.384	0.151	-0.116	0.258	0.091	11.708
56	-0.463	0.187	0.076	0.237	-0.037	34.633
57	-0.263	-0.234	-0.003	0.176	0.324	108.749
58	-0.304	-0.196	0.052	0.216	0.232	138.776
59	-0.296	-0.204	0.073	0.183	0.243	356.685
60	-0.267	0.256	0.244	-0.149	-0.084	21.194
61	-0.500	0.195	0.032	0.267	0.007	16.917
62	-0.276	-0.224	0.016	0.218	0.267	368.469
63	-0.322	-0.178	0.007	0.173	0.321	101.661
64	-0.281	-0.219	0.011	0.225	0.264	50.511
65	-0.500	0.118	0.128	0.080	0.174	8.389
66	-0.306	-0.194	0.029	0.209	0.262	208.340
67	-0.460	-0.040	0.066	0.197	0.237	22.598
68	-0.375	-0.125	0.036	0.199	0.265	27.917
69	-0.269	-0.193	-0.038	0.214	0.286	106.540
70	-0.340	-0.160	0.055	0.268	0.177	39.608
71	-0.333	-0.141	-0.027	0.100	0.400	34.882
72	-0.286	-0.198	-0.017	0.216	0.284	79.611
73	-0.261	-0.239	0.005	0.216	0.279	370.943
74	-0.350	-0.150	0.002	0.228	0.270	57.080
75	-0.281	-0.219	0.025	0.217	0.258	203.470
76	-0.259	-0.126	-0.115	0.228	0.272	33.726
77	-0.317	-0.181	-0.003	0.222	0.278	336.413
78	-0.392	-0.076	-0.032	0.187	0.313	45.098
79	-0.280	-0.195	-0.025	0.203	0.297	346.395
80	-0.291	-0.194	-0.014	0.226	0.274	245.295
81	-0.300	-0.200	0.094	0.201	0.205	52.442
82	-0.217	-0.283	0.118	0.106	0.276	156.597
83	-0.253	-0.247	0.075	0.167	0.258	58.472
84	-0.338	0.183	0.201	-0.162	0.116	5.194
85	-0.313	-0.157	-0.029	0.214	0.286	214.329
86	-0.500	0.017	0.364	0.101	0.017	6.608
87	-0.408	0.363	-0.053	-0.040	0.137	28.754
88	-0.380	-0.120	0.025	0.094	0.381	47.214
89	-0.312	-0.188	0.023	0.230	0.247	144.702
90	-0.321	-0.178	-0.001	0.225	0.275	335.485
91	-0.303	-0.046	-0.151	0.161	0.339	49.618
92	-0.266	-0.226	-0.008	0.216	0.284	108.692
93	-0.354	-0.146	0.152	0.190	0.158	19.514
94	-0.332	-0.134	-0.033	0.226	0.274	97.221
95	-0.399	-0.057	-0.045	0.223	0.277	57.660
96	-0.239	-0.261	0.006	0.212	0.282	77.750
97	-0.271	-0.229	0.070	0.203	0.227	72.345
98	-0.283	-0.113	-0.104	0.136	0.364	67.646
99	-0.384	0.046	-0.116	0.105	0.348	188.230
100	-0.450	-0.050	0.005	0.287	0.208	37.592
101	-0.261	-0.239	0.028	0.201	0.271	41.622
102	-0.338	-0.162	0.022	0.228	0.250	82.687
103	-0.389	-0.086	-0.025	0.211	0.289	34.940
104	-0.293	-0.207	0.022	0.153	0.325	91.462
105	-0.052	0.393	0.107	-0.159	-0.289	19.821
106	-0.439	-0.061	0.004	0.243	0.253	68.279
107	0.227	0.273	-0.068	-0.144	-0.288	32.069

Table B.15: The value of the likelihood ratio statistic and the corresponding p-value for subjects of control treatment.

ID	AM		FIV		ONE		BIN		POS		NEG		MED	
1	16.444	(0.001)	15.042	(0.002)	19.281	(0.000)	19.036	(0.000)	14.457	(0.002)	18.515	(0.000)	17.863	(0.000)
2	8.381	(0.039)	14.977	(0.002)	14.410	(0.002)	5.125	(0.163)	11.447	(0.010)	10.110	(0.018)	13.330	(0.004)
3	22.009	(0.000)	2.666	(0.446)	22.395	(0.000)	22.621	(0.000)	23.307	(0.000)	22.246	(0.000)	15.259	(0.002)
4	1.773	(0.621)	1.677	(0.642)	2.521	(0.471)	2.607	(0.456)	2.385	(0.496)	2.484	(0.478)	2.501	(0.475)
5	10.382	(0.016)	15.108	(0.002)	9.513	(0.023)	8.473	(0.037)	16.199	(0.001)	11.180	(0.011)	15.610	(0.001)
6	7.679	(0.053)	7.784	(0.051)	7.568	(0.056)	4.403	(0.221)	0.704	(0.872)	7.679	(0.053)	6.314	(0.097)
7	4.396	(0.222)	6.268	(0.099)	4.202	(0.240)	4.961	(0.175)	5.434	(0.143)	6.340	(0.096)	5.675	(0.129)
8	0.675	(0.879)	17.776	(0.000)	10.290	(0.016)	11.308	(0.010)	17.182	(0.001)	9.656	(0.022)	17.927	(0.000)
9	3.610	(0.307)	5.593	(0.133)	3.961	(0.266)	0.815	(0.846)	6.305	(0.098)	2.449	(0.485)	6.223	(0.101)
10														
11	17.074	(0.001)	31.408	(0.000)	27.284	(0.000)	18.715	(0.000)	22.918	(0.000)	27.955	(0.000)	31.228	(0.000)
12	13.247	(0.004)	14.548	(0.002)	7.628	(0.054)	13.555	(0.004)	21.444	(0.000)	8.216	(0.042)	15.122	(0.002)
13	10.563	(0.014)	15.412	(0.001)	13.537	(0.004)	1.735	(0.629)	14.282	(0.003)	8.702	(0.034)	17.044	(0.001)
14	9.225	(0.026)	17.815	(0.000)	11.541	(0.009)	9.102	(0.028)	17.881	(0.000)	13.156	(0.004)	18.451	(0.000)
15	18.972	(0.000)	32.315	(0.000)	29.698	(0.000)	16.151	(0.001)	26.185	(0.000)	27.316	(0.000)	32.082	(0.000)
16	6.918	(0.075)	9.203	(0.027)	8.540	(0.036)	4.929	(0.177)	4.255	(0.235)	9.230	(0.026)	9.411	(0.024)
17	1.535	(0.674)	14.998	(0.002)	9.021	(0.029)	7.442	(0.059)	13.919	(0.003)	8.280	(0.041)	15.104	(0.002)
18	2.209	(0.530)	20.689	(0.000)	10.003	(0.019)	13.272	(0.004)	19.901	(0.000)	13.352	(0.004)	20.270	(0.000)
19	18.698	(0.000)	21.554	(0.000)	22.388	(0.000)	12.156	(0.007)	9.411	(0.024)	21.522	(0.000)	18.274	(0.000)
20	14.311	(0.003)	33.684	(0.000)	16.321	(0.001)	24.658	(0.000)	34.386	(0.000)	15.220	(0.002)	32.673	(0.000)
21	2.256	(0.521)	11.549	(0.009)	7.559	(0.056)	4.054	(0.256)	9.673	(0.022)	6.397	(0.094)	11.548	(0.009)
22	1.885	(0.597)	1.583	(0.663)	2.076	(0.557)	0.479	(0.923)	1.274	(0.735)	1.722	(0.632)	2.332	(0.506)
23	1.212	(0.750)	13.542	(0.004)	7.840	(0.049)	7.545	(0.056)	11.730	(0.008)	8.675	(0.034)	13.522	(0.004)
24	1.717	(0.633)	15.239	(0.002)	8.013	(0.046)	8.264	(0.041)	14.536	(0.002)	8.129	(0.043)	14.968	(0.002)
25	22.443	(0.000)	42.581	(0.000)	17.760	(0.000)	36.939	(0.000)	42.931	(0.000)	31.757	(0.000)	39.892	(0.000)
26	17.838	(0.000)	32.117	(0.000)	24.491	(0.000)	17.028	(0.001)	31.413	(0.000)	10.479	(0.015)	32.681	(0.000)
27	7.932	(0.047)	8.731	(0.033)	7.161	(0.067)	7.525	(0.057)	10.925	(0.012)	9.851	(0.020)	8.665	(0.034)
28	1.721	(0.632)	14.660	(0.002)	9.395	(0.024)	8.918	(0.030)	13.927	(0.003)	10.188	(0.017)	14.879	(0.002)
29	5.516	(0.138)	14.685	(0.002)	7.443	(0.059)	12.692	(0.005)	14.310	(0.003)	6.065	(0.108)	14.516	(0.002)
30	6.021	(0.111)	7.189	(0.066)	4.841	(0.184)	6.315	(0.097)	8.172	(0.043)	0.510	(0.917)	7.321	(0.062)
31	9.726	(0.021)	7.215	(0.065)	10.721	(0.013)	10.575	(0.014)	6.580	(0.087)	10.485	(0.015)	10.637	(0.014)
32	5.336	(0.149)	7.912	(0.048)	6.831	(0.077)	0.519	(0.915)	7.364	(0.061)	3.235	(0.357)	8.768	(0.033)
33	20.526	(0.000)	27.884	(0.000)	26.031	(0.000)	8.929	(0.030)	14.736	(0.002)	25.044	(0.000)	28.114	(0.000)
34	2.163	(0.539)	6.451	(0.092)	0.940	(0.816)	5.956	(0.114)	6.383	(0.094)	4.757	(0.190)	4.755	(0.191)
35	15.313	(0.002)	21.673	(0.000)	19.126	(0.000)	5.401	(0.145)	18.931	(0.000)	15.097	(0.002)	23.346	(0.000)
36	6.761	(0.080)	21.969	(0.000)	19.841	(0.000)	16.645	(0.001)	20.466	(0.000)	19.154	(0.000)	22.871	(0.000)
37	12.905	(0.005)	26.224	(0.000)	18.702	(0.000)	11.868	(0.008)	25.524	(0.000)	7.738	(0.052)	26.556	(0.000)
38	9.276	(0.026)	16.541	(0.001)	12.701	(0.005)	3.213	(0.360)	15.252	(0.002)	5.206	(0.157)	17.294	(0.001)
39	9.518	(0.023)	15.030	(0.002)	12.926	(0.005)	1.629	(0.653)	12.931	(0.005)	8.744	(0.033)	16.221	(0.001)
40	13.920	(0.003)	29.595	(0.000)	12.994	(0.005)	22.371	(0.000)	30.974	(0.000)	21.223	(0.000)	28.054	(0.000)
41	1.680	(0.641)	13.532	(0.004)	5.784	(0.123)	7.000	(0.072)	12.704	(0.005)	7.656	(0.054)	13.088	(0.004)
42	26.773	(0.000)	29.432	(0.000)	24.263	(0.000)	21.469	(0.000)	35.508	(0.000)	16.520	(0.001)	32.777	(0.000)
43	16.507	(0.001)	7.114	(0.068)	13.854	(0.003)	16.507	(0.001)	16.512	(0.001)	16.508	(0.001)	5.050	(0.168)
44	6.320	(0.097)	9.997	(0.019)	7.217	(0.065)	3.818	(0.282)	9.434	(0.024)	7.508	(0.057)	10.511	(0.015)
45	8.246	(0.041)	24.370	(0.000)	17.567	(0.001)	11.454	(0.010)	21.434	(0.000)	16.021	(0.001)	24.366	(0.000)
46	4.146	(0.246)	18.811	(0.000)	6.021	(0.111)	14.250	(0.003)	18.797	(0.000)	6.394	(0.094)	17.774	(0.000)
47	4.199	(0.241)	6.383	(0.094)	5.640	(0.130)	3.979	(0.264)	2.644	(0.450)	6.064	(0.109)	6.382	(0.094)
48	4.067	(0.254)	22.482	(0.000)	15.771	(0.001)	21.205	(0.000)	23.818	(0.000)	20.059	(0.000)	23.864	(0.000)
49	5.958	(0.114)	14.287	(0.003)	9.876	(0.020)	5.246	(0.155)	11.099	(0.011)	9.734	(0.021)	14.576	(0.002)
50	0.863	(0.834)	19.106	(0.000)	12.200	(0.007)	15.074	(0.002)	19.509	(0.000)	12.770	(0.005)	19.830	(0.000)
51	7.175	(0.067)	10.473	(0.015)	13.802	(0.003)	13.614	(0.003)	14.795	(0.002)	13.031	(0.005)	13.374	(0.004)
52	7.763	(0.051)	17.960	(0.000)	17.880	(0.000)	19.176	(0.000)	21.639	(0.000)	16.842	(0.001)	20.564	(0.000)
53	3.779	(0.286)	19.871	(0.000)	4.597	(0.204)	14.350	(0.002)	19.844	(0.000)	9.730	(0.021)	18.543	(0.000)

Table B.16: The value of the likelihood ratio statistic and the corresponding p-value for subjects of information treatment.

ID	AM		FIV		ONE		BIN		POS		NEG		MED	
54	9.976	(0.019)	25.460	(0.000)	23.050	(0.000)	18.556	(0.000)	23.867	(0.000)	23.297	(0.000)	26.601	(0.000)
55	2.276	(0.517)	0.782	(0.854)	1.581	(0.664)	1.091	(0.779)	2.366	(0.500)	1.410	(0.703)	1.710	(0.635)
56	12.388	(0.006)	10.239	(0.017)	4.540	(0.209)	13.451	(0.004)	17.016	(0.001)	7.867	(0.049)	8.158	(0.043)
57	7.461	(0.059)	22.861	(0.000)	19.284	(0.000)	14.079	(0.003)	21.222	(0.000)	17.305	(0.001)	23.378	(0.000)
58	14.880	(0.002)	21.982	(0.000)	17.267	(0.001)	9.332	(0.025)	24.057	(0.000)	4.698	(0.195)	23.710	(0.000)
59	14.755	(0.002)	32.630	(0.000)	23.943	(0.000)	19.138	(0.000)	31.823	(0.000)	10.335	(0.016)	32.652	(0.000)
60	13.587	(0.004)	17.865	(0.000)	5.332	(0.149)	18.194	(0.000)	15.379	(0.002)	15.239	(0.002)	11.179	(0.011)
61	12.425	(0.006)	8.830	(0.032)	7.514	(0.057)	11.985	(0.007)	15.855	(0.001)	11.477	(0.009)	9.542	(0.023)
62														
63	1.300	(0.729)	26.239	(0.000)	17.091	(0.001)	18.330	(0.000)	25.393	(0.000)	17.191	(0.001)	26.628	(0.000)
64	13.737	(0.003)	15.777	(0.001)	16.060	(0.001)	2.547	(0.467)	8.872	(0.031)	13.475	(0.004)	17.682	(0.001)
65	5.868	(0.118)	7.551	(0.056)	4.621	(0.202)	7.115	(0.068)	5.826	(0.120)	2.918	(0.405)	6.974	(0.073)
66	11.736	(0.008)	27.290	(0.000)	19.355	(0.000)	13.242	(0.004)	25.975	(0.000)	10.754	(0.013)	27.383	(0.000)
67	1.972	(0.578)	8.936	(0.030)	4.476	(0.214)	5.850	(0.119)	8.912	(0.030)	1.699	(0.637)	8.714	(0.033)
68	2.602	(0.457)	11.729	(0.008)	4.130	(0.248)	5.180	(0.159)	11.645	(0.009)	4.436	(0.218)	11.088	(0.011)
69	13.285	(0.004)	22.732	(0.000)	20.888	(0.000)	9.071	(0.028)	15.809	(0.001)	19.347	(0.000)	22.403	(0.000)
70	14.533	(0.002)	11.967	(0.007)	14.691	(0.002)	4.444	(0.217)	15.835	(0.001)	7.654	(0.054)	17.607	(0.001)
71	3.409	(0.333)	8.016	(0.046)	8.616	(0.035)	10.140	(0.017)	10.556	(0.014)	9.680	(0.021)	10.289	(0.016)
72	5.452	(0.142)	15.091	(0.002)	10.330	(0.016)	5.024	(0.170)	12.835	(0.005)	9.963	(0.019)	15.003	(0.002)
73	21.458	(0.000)	30.562	(0.000)	28.001	(0.000)	9.227	(0.026)	23.937	(0.000)	22.030	(0.000)	30.484	(0.000)
74	5.533	(0.137)	15.423	(0.001)	10.733	(0.013)	4.152	(0.245)	13.640	(0.003)	7.640	(0.054)	15.736	(0.001)
75	19.440	(0.000)	26.756	(0.000)	23.508	(0.000)	7.726	(0.052)	24.573	(0.000)	14.011	(0.003)	28.210	(0.000)
76	4.304	(0.230)	6.052	(0.109)	5.179	(0.159)	0.283	(0.963)	5.319	(0.150)	3.183	(0.364)	6.841	(0.077)
77	16.731	(0.001)	35.571	(0.000)	28.810	(0.000)	20.366	(0.000)	32.299	(0.000)	25.335	(0.000)	35.560	(0.000)
78	4.261	(0.235)	14.665	(0.002)	10.364	(0.016)	8.996	(0.029)	11.578	(0.009)	12.195	(0.007)	14.671	(0.002)
79														
80	18.784	(0.000)	28.814	(0.000)	26.064	(0.000)	8.640	(0.034)	20.658	(0.000)	23.410	(0.000)	28.781	(0.000)
81	10.744	(0.013)	17.766	(0.000)	12.107	(0.007)	6.579	(0.087)	18.824	(0.000)	2.736	(0.434)	19.044	(0.000)
82	17.377	(0.001)	28.119	(0.000)	28.004	(0.000)	16.917	(0.001)	23.951	(0.000)	23.125	(0.000)	25.760	(0.000)
83	5.259	(0.154)	13.129	(0.004)	10.199	(0.017)	4.388	(0.222)	12.156	(0.007)	3.816	(0.282)	12.924	(0.005)
84	0.520	(0.914)	2.051	(0.562)	0.704	(0.872)	1.673	(0.643)	1.995	(0.573)	1.683	(0.641)	1.542	(0.673)
85	10.759	(0.013)	28.339	(0.000)	22.508	(0.000)	16.586	(0.001)	25.085	(0.000)	22.423	(0.000)	28.489	(0.000)
86	2.023	(0.568)	2.969	(0.396)	0.561	(0.905)	2.624	(0.453)	3.412	(0.332)	0.792	(0.851)	2.230	(0.526)
87	3.022	(0.388)	7.731	(0.052)	5.580	(0.134)	8.699	(0.034)	8.732	(0.033)	8.171	(0.043)	8.621	(0.035)
88	2.909	(0.406)	15.679	(0.001)	11.429	(0.010)	16.081	(0.001)	17.578	(0.001)	13.750	(0.003)	17.580	(0.001)
89	23.437	(0.000)	27.319	(0.000)	26.187	(0.000)	8.955	(0.030)	27.939	(0.000)	15.592	(0.001)	30.739	(0.000)
90	15.571	(0.001)	33.438	(0.000)	26.527	(0.000)	18.111	(0.000)	30.228	(0.000)	23.068	(0.000)	33.452	(0.000)
91	5.191	(0.158)	8.752	(0.033)	10.408	(0.015)	12.213	(0.007)	12.388	(0.006)	12.097	(0.007)	12.231	(0.007)
92	11.729	(0.008)	20.762	(0.000)	17.797	(0.000)	5.110	(0.164)	15.482	(0.001)	13.910	(0.003)	20.764	(0.000)
93	12.765	(0.005)	10.291	(0.016)	13.506	(0.004)	1.624	(0.654)	12.276	(0.006)	6.065	(0.108)	15.047	(0.002)
94	11.966	(0.007)	23.263	(0.000)	18.182	(0.000)	10.911	(0.012)	17.804	(0.000)	18.397	(0.000)	23.431	(0.000)
95	3.814	(0.282)	15.927	(0.001)	10.590	(0.014)	6.144	(0.105)	13.903	(0.003)	9.479	(0.024)	15.970	(0.001)
96	13.914	(0.003)	19.013	(0.000)	18.334	(0.000)	5.791	(0.122)	8.126	(0.043)	16.890	(0.001)	18.276	(0.000)
97	9.957	(0.019)	12.909	(0.005)	8.697	(0.034)	5.444	(0.142)	15.452	(0.001)	3.514	(0.319)	13.933	(0.003)
98	6.761	(0.080)	14.083	(0.003)	12.455	(0.006)	14.157	(0.003)	14.362	(0.002)	15.035	(0.002)	15.422	(0.001)
99	15.673	(0.001)	31.846	(0.000)	28.912	(0.000)	28.866	(0.000)	30.472	(0.000)	31.635	(0.000)	33.316	(0.000)
100	8.707	(0.033)	10.883	(0.012)	8.542	(0.036)	6.614	(0.085)	10.973	(0.012)	10.246	(0.017)	11.751	(0.008)
101	11.368	(0.010)	12.818	(0.005)	13.100	(0.004)	0.354	(0.950)	10.488	(0.015)	7.992	(0.046)	14.303	(0.003)
102	26.613	(0.000)	32.009	(0.000)	28.888	(0.000)	18.296	(0.000)	34.497	(0.000)	0.813	(0.846)	34.662	(0.000)
103	1.634	(0.652)	10.168	(0.017)	6.519	(0.089)	3.987	(0.263)	8.938	(0.030)	5.524	(0.137)	10.065	(0.018)
104	2.037	(0.565)	23.169	(0.000)	16.185	(0.001)	16.520	(0.001)	22.974	(0.000)	14.590	(0.002)	23.661	(0.000)
105	3.994	(0.262)	2.898	(0.408)	1.633	(0.652)	4.396	(0.222)	4.683	(0.197)	4.178	(0.243)	0.581	(0.901)
106	5.983	(0.112)	17.823	(0.000)	9.913	(0.019)	8.938	(0.030)	18.121	(0.000)	9.278	(0.026)	17.739	(0.000)
107	1.364	(0.714)	5.333	(0.149)	3.587	(0.310)	2.986	(0.394)	4.141	(0.247)	4.178	(0.243)	5.414	(0.144)

Table B.17: Win/tie/loss statistics for estimates in comparison with reference functions, separately for control (top) and information (bottom) treatment.

	AM	FIV	ONE	BIN	POS	NEG	MED
Estimates (CT)	139 / 434 / 63	377 / 194 / 65	248 / 293 / 95	260 / 285 / 91	293 / 281 / 62	249 / 288 / 99	306 / 270 / 60
Estimates (IT)	172 / 400 / 76	404 / 159 / 85	249 / 316 / 83	252 / 296 / 100	284 / 296 / 68	259 / 295 / 94	320 / 256 / 72

Table B.18: Average likelihood ratio statistic for the different reference aggregation functions (separately for control and information treatment).

	AM	FIV	ONE	BIN	POS	NEG	MED
CT	8.838	16.437	12.346	10.867	15.919	11.863	16.885
IT	9.562	17.352	13.971	9.598	16.350	11.452	17.845

Appendix C.

Figure C.11 depicts the dendrogram of hierarchical weighted average-linkage clustering.

Including the estimated weights of the reference functions (cf. Table A.12), we use the number of clusters that separate these reference functions. Hence, in information treatment we have 14 clusters. In control treatment, reference functions are already separated with 10 clusters.[14] These raw clusters are provided in Tables C.19 and C.20. Refining the clusters, we group clusters that focus on the 1-star category. Note that the resulting clusters are phrased in accordance with the reference heuristics they contain. Hence, category weights differ across treatments although they have the same name (cf. Table C.21). In the information treatment, cluster (d) does not contain a reference function and has been phrased as a combination of the two heuristics AM and POS.

[14] We increased the number of groups and found clusters to be rather robust.

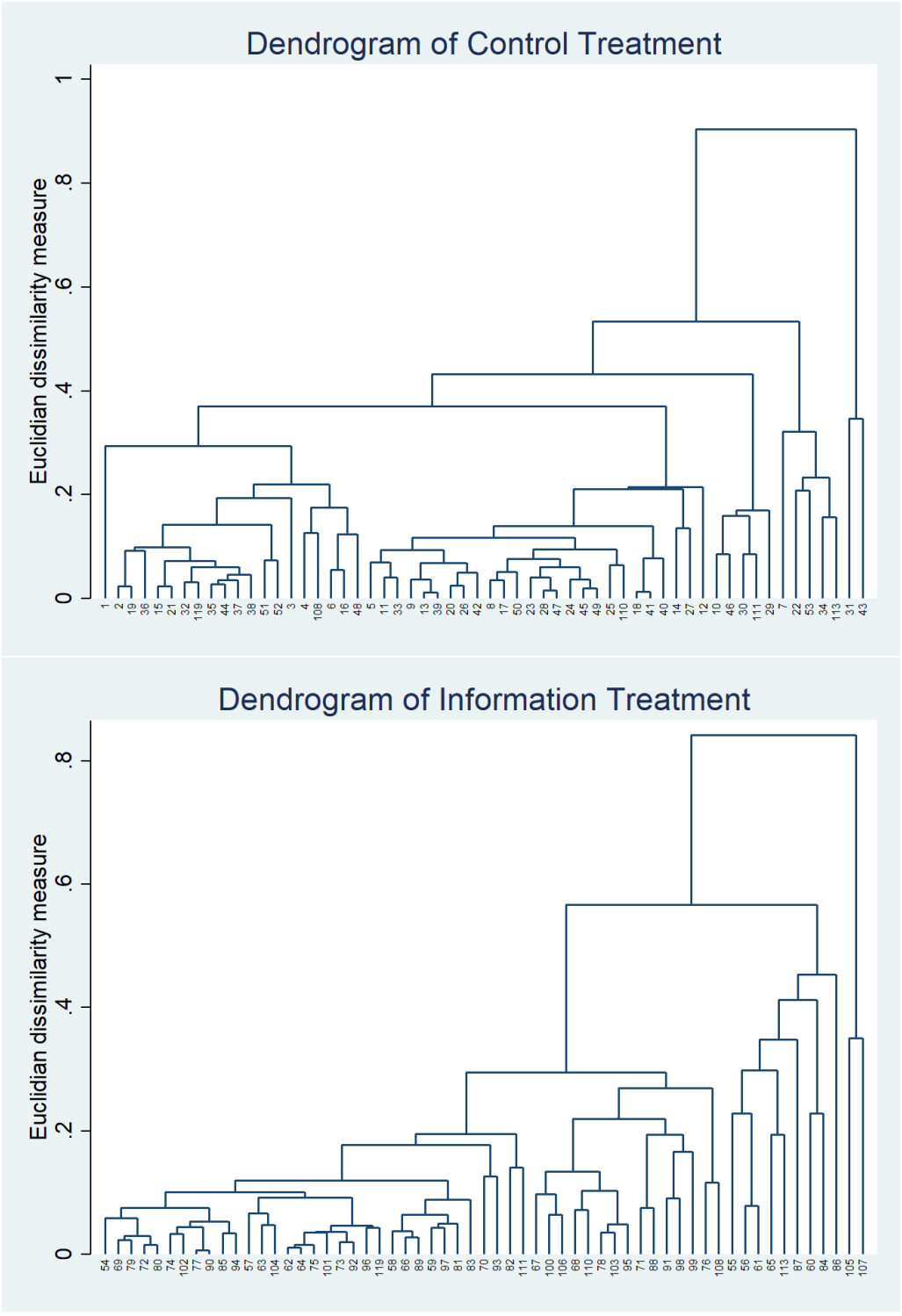


Figure C.11: Dendrogram for control treatment (top) and information treatment (bottom) calculated with hierarchical weighted average-linkage clustering. Reference functions have the IDs 108 (POS), 110 (AM), 111 (NEG), 113 (ONE), and 119 (BIN). FIV and MED have been discarded in the preceding analysis.

Table C.19: Resulting clusters in information treatment. According to amount, subjects of cluster (a) systematically overweight moderate categories w_2 and w_4 at the expense of extreme categories w_1 and w_5 (**BIN**). Subjects of cluster (c) show behavior that can be explained by **AM**. Cluster (d) weights categories w_2 and w_4 similarly, resulting in (**AM-POS**). Groups (g), (h), (i), and (l) have focus on the minimizing of 1-star ratings (**ONE**) and are hence collapsed to one group in the subsequent analysis. We find no subjects whose behavior is associated with FIV or MED. Additionally, POS (e) resp. NEG (b) is only applied by a single subject. Category weights of groups (f), (j), (k), (m), and (n) are not identifiable.

Cluster		a (BIN)	b (NEG)	c (AM)	d (AM-POS)	e (POS)	f	g (ONE)	h (ONE)	i (ONE)	j	k	l (ONE)	m	n	
Contain Reference		BIN*	NEG*	AM*		POS*			ONE*							
#		30	1	7	5	1	1	2	1	1	1	1	1	1	1	54
w1	mean	-0.295	-0.217	-0.415	-0.337	-0.259	-0.384	-0.481	-0.5	-0.408	-0.267	-0.338	-0.5	-0.052	0.227	-0.317
	std.dev.	0.03		0.034	0.045			0.026								0.11
	median	-0.292		-0.399	-0.333			-0.481								-0.309
w2	mean	-0.197	-0.283	-0.071	-0.075	-0.126	0.151	0.191	0.118	0.363	0.256	0.183	0.017	0.393	0.273	-0.094
	std.dev.	0.032		0.028	0.076			0.005								0.167
	median	-0.196		-0.061	-0.113			0.191								-0.159
w3	mean	0.018	0.118	0.001	-0.074	-0.115	-0.116	0.054	0.128	-0.053	0.244	0.201	0.364	0.107	-0.068	0.02
	std.dev.	0.043		0.039	0.072			0.032								0.088
	median	0.009		0.004	-0.104			0.054								0.006
w4	mean	0.209	0.106	0.221	0.119	0.228	0.258	0.252	0.08	-0.04	-0.149	-0.162	0.101	-0.159	-0.144	0.167
	std.dev.	0.023		0.034	0.028			0.022								0.107
	median	0.215		0.211	0.105			0.252								0.206
w5	mean	0.266	0.276	0.263	0.366	0.272	0.091	-0.015	0.174	0.137	-0.084	0.116	0.017	-0.289	-0.288	0.223
	std.dev.	0.038		0.034	0.024			0.031								0.14
	median	0.272		0.265	0.364			-0.015								0.268
alpha	mean	154.03	156.6	42.01	77.52	33.73	11.71	25.77	8.39	28.75	21.19	5.19	6.61	19.82	32.07	105.15
	std.dev.	114.48		16.28	62.99			12.53								105.04
	median	106.04		37.59	49.62			25.77								63.06

Table C.20: Resulting clusters in control treatment. Cluster (d) is most in line with the arithmetic mean (**AM**). Cluster (b) show rather binary weighting (**BIN**). Cluster (c) show specifications that are mostly in accordance with **POS**. Weighting w_3 strongly positive and not discriminate between non-negative categories, subjects in cluster (e) show **NEG**-alike behavior. Weightings of cluster (f), (g), and (h) show strong negative weight on w_1, suggesting **ONE**-weighting. There is no evidence for the employment of FIV or MED. Three outliers (a), (i), and (j) have non-interpretable category weights.

Cluster		a	b (BIN)	c (POS)	d (AM)	e (NEG)	f (ONE)	g (ONE)	h (ONE)	i	j	
Contain Reference			BIN*	POS*	AM*	NEG*			ONE*			
#		1	13	4	25	4	1	2	1	1	1	53
w1	mean	-0.286	-0.265	-0.276	-0.354	-0.331	-0.448	-0.5	-0.5	0.221	0.068	-0.315
	std.dev.		0.031	0.046	0.049	0.03		0				0.117
	median		-0.272	-0.274	-0.335	-0.332		-0.5				-0.321
w2	mean	0.045	-0.223	-0.097	-0.121	-0.169	0.407	0.116	0.196	0.159	0.432	-0.104
	std.dev.		0.027	0.053	0.065	0.03		0.098				0.15
	median		-0.213	-0.076	-0.146	-0.168		0.116				-0.152
w3	mean	-0.214	0.006	-0.126	-0.011	0.204	-0.052	0.063	0.05	0.12	-0.015	0.002
	std.dev.		0.041	0.033	0.045	0.053		0.077				0.087
	median		0.003	-0.127	-0.009	0.181		0.063				0.002
w4	mean	0.067	0.173	0.184	0.218	0.096	0.03	0.066	0.108	-0.14	-0.086	0.169
	std.dev.		0.056	0.032	0.036	0.065		0.065				0.086
	median		0.197	0.185	0.22	0.09		0.066				0.195
w5	mean	0.387	0.309	0.316	0.269	0.2	0.063	0.256	0.145	-0.36	-0.399	0.248
	std.dev.		0.056	0.032	0.041	0.056		0.044				0.141
	median		0.303	0.315	0.275	0.199		0.256				0.275
alpha	mean	37.35	94.36	39.73	108.9	108.7	11.67	21.33	12.47	31.23	40.96	89.04
	std.dev.		81.19	35.29	99.52	140.4		20.72				91.04
	median		58.13	28.65	62.74	49.39		21.33				56.32

Table C.21: Mean weights (and std. dev.) of identified clusters.

Treatment	CT					IT			
Cluster	AM	BIN	POS	NEG	ONE	BIN	AM	AM/POS	ONE
#	25	13	4	4	4	30	7	5	5
w_1	-0.354 (0.049)	-0.265 (0.031)	-0.276 (0.046)	-0.331 (0.030)	-0.487 (0.026)	-0.295 (0.030)	-0.415 (0.034)	-0.337 (0.045)	-0.474 (0.040)
w_2	-0.121 (0.065)	-0.223 (0.027)	-0.097 (0.052)	-0.169 (0.030)	0.209 (0.149)	-0.197 (0.032)	-0.071 (0.028)	-0.075 (0.076)	0.176 (0.127)
w_3	-0.011 (0.045)	0.006 (0.041)	-0.126 (0.033)	0.204 (0.053)	0.031 (0.071)	0.018 (0.043)	0.001 (0.039)	-0.074 (0.072)	0.110 (0.157)
w_4	0.218 (0.036)	0.173 (0.056)	0.184 (0.032)	0.096 (0.065)	0.067 (0.049)	0.209 (0.023)	0.221 (0.034)	0.119 (0.028)	0.129 (0.125)
w_5	0.269 (0.041)	0.309 (0.056)	0.316 (0.032)	0.200 (0.056)	0.180 (0.097)	0.266 (0.038)	0.263 (0.034)	0.366 (0.024)	0.059 (0.091)
α	108.90 (99.52)	94.36 (81.19)	39.73 (35.29)	108.70 (140.40)	16.70 (13.11)	154.03 (114.48)	42.01 (16.28)	77.52 (62.99)	19.06 (12.35)

Appendix D.

Table D.22: Pairwise comparison between groups with regard to the congruency with the arithmetic mean. Values in group cells correspond to mean values of the α^{Mean} parameter. Std. dev. are presented in parentheses. z- and p-values calculated with MWU-test.

		z	p
Male (n=37) 31.91 (22.11)	Female (n=70) 38.58 (19.33)	-1.919	0.055
Older (n=45) 36.90 (22.00)	Younger (n=62) 35.82 (19.48)	-0.133	0.8946
Non-freshmen (n=45) 35.43 (18.61)	Freshmen (n=62) 36.89 (21.87)	0.233	0.8154
Economics (n=39) 36.27 (20.05)	Education (n=44) 36.31 (21.26)	0.274	0.8943
Economics (n=39) 36.27 (20.05)	Engineering (n= 11) 30.36 (22.74)	0.902	0.3672
Economics (n=39) 36.27 (20.05)	Humanities (n=11) 42.95 (18.60)	-0.995	0.3196
Education (n=44) 36.31 (21.26)	Engineering (n= 11) 30.36 (22.74)	0.968	0.3331
Education (n=44) 36.31 (21.26)	Humanities (n=11) 42.95 (18.60)	-1.22	0.2223
Engineering (n= 11) 30.36 (22.74)	Humanities (n=11) 42.95 (18.60)	-1.477	0.1396

Table D.23: Demographics of identified clusters. z-values and p-values of χ^2-tests are reported.

Treatment	Cluster	n	Male	Female	Older	Younger	Wise	Freshmen	ECO	EDU	ENG	HUM
CT	AM	25	5	20	12	13	15	10	10	12	1	2
	BIN	13	5	8	4	9	8	5	5	2	2	4
	POS	4	2	2	2	2	3	1	2	1	0	0
	NEG	4	2	2	1	3	1	3	2	2	0	0
	ONE	4	1	3	2	2	1	3	0	4	0	0
			z=3.2051, p=0.524		z=1.7273, p=0.786		z=4.0303, p=0.402		z=15.9322, p=0.194			
IT	BIN	30	13	17	14	16	9	21	11	14	3	1
	AM	7	2	5	2	5	1	6	0	5	1	1
	AM/POS	5	2	3	4	1	1	4	3	0	0	2
	ONE	5	2	3	1	4	1	4	2	1	1	1
			z=0.5145, p=0.916		z=4.5385, p=0.209		z=0.9416, p=0.815		z=14.5308, p=0.105			

Table D.24: No difference between clusters with regard to online shopping.

Treatment	Cluster	n	Experienced	Non-Experienced
CT	AM	25	6	19
	BIN	13	4	9
	POS	4	1	3
	NEG	4	1	3
	ONE	4	1	3
			z=0.2119, p=0.995	
IT	BIN	30	7	23
	AM	7	2	5
	AM/POS	5	1	4
	ONE	5	2	3
			z=0.7412, p=0.863	

Table D.25: No difference between clusters with regard to general risk attitudes.

Treatment	Cluster	n	More Risk-seeking	Less Risk-seeking
CT	AM	25	8	17
	BIN	13	7	6
	POS	4	3	1
	NEG	4	1	3
	ONE	4	2	2
			z=4.1430, p=0.387	
IT	BIN	30	16	14
	AM	7	4	3
	AM/POS	5	4	1
	ONE	5	3	2
			z=1.2616, p=0.738	