Decision

Recommender Systems: Friend (of Choice) or Foe? A Large-Scale Field Experiment in Online Shopping Platforms

Nurit Nobel

Online First Publication, May 16, 2024. https://dx.doi.org/10.1037/dec0000236

AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

https://doi.org/10.1037/dec0000236

Recommender Systems: Friend (of Choice) or Foe? A Large-Scale Field Experiment in Online Shopping Platforms

Nurit Nobel

The Sustainability Transparency Accountability Research Lab, Harvard University
Department of Marketing and Strategy, Stockholm School of Economics

Recommender system (RS) algorithms are increasingly involved in online decision making, helping consumers quickly screen large assortments by presenting the most appealing products first. However, while purporting to facilitate choice, RS algorithms might be inadvertently hindering it due to the highly attractive choice sets they produce which might contribute to choice overload. In a field experiment conducted with 23,165 consumers in two online retailers, I find no indication that a considerably altered RS algorithm that presented consumers with only the most appealing product results followed by results of diminished attractiveness hurt choice process or outcome. The findings, coupled with a follow-up equivalence analysis, suggest that a small set of highly attractive products followed by many less appealing options may work as effectively as a large set of only attractive options in terms of choice process measures. This research proposes that studying consumer psychology phenomena in the unique setting provided by RS algorithms can lead to advances in both scientific theory and algorithm design.

Keywords: choice overload, recommender systems, consumer behavior, algorithms, field experiment

In the recent years, the ubiquity of algorithms in online choice situations has increased. These days, algorithms make recommendations to people regarding which movies to see, books to read, products to buy and restaurants to visit (Köcher et al., 2019). These algorithms, generally referred to as recommender systems (RS), guide individuals through the exploration of large amounts of online content by increasing the salience of items these individuals are more likely to find appealing (Jannach et al., 2010). RS algorithms rely on machine learning (ML) to generate relevant and personalized product recommendations based on an analysis of both explicit (e.g., product ratings) and implicit (e.g., browsing patterns) past user behavior (Marchand & Marx, 2020).

By origin, RS algorithms are designed to support consumers in choice situations characterized by seemingly endless assortments offered by online retailers (J. Ma, 2016). RS algorithms

Nurit Nobel https://orcid.org/0000-0001-7571-4380

This work was supported by the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbanken Forskningsstiftelser, Grants P18 0073 and P21 0091), as well as the Knut and Alice Wallenberg Foundation and the Marianne and Marcus Wallenberg Foundation granted to Anna Dreber. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the article.

The author thanks Patric Andersson, Anna Dreber, Shai Davidai, Magnus Johannesson, Thomas Lavelle, Rickard Sandberg, and seminar participants at Stockholm School of Economics for helpful feedback and suggestions. The author is also grateful to Tina Emambakhsh and Almog Simchon who provided invaluble research assistance and advice. Finally, the author thanks the Findify team for their collaboration in conducting this field experiment.

Findify, the industry partner in this study, was provided an opportunity to review article content for factual inaccuracies, but the author maintained full editorial control. The findings and conclusions expressed are solely those of the author and do not represent the views of Harvard University, Stockholm School of Economics, or Findify. Earlier versions of this work were previously presented in research conferences and posted on the online repository Social Science Research Network.

Nurit Nobel played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Nurit Nobel, The Sustainability Transparency Accountability Research Lab, Harvard University, 1727 Cambridge Street, Cambridge, MA 02138, United States. Email: nnobel@fas.harvard.edu

reduce information overload by presenting consumers with the most relevant alternatives, thus ridding them of the need to screen through extensive product lists themselves (Häubl & Trifts, 2000; Köcher et al., 2019). However, RS might inadvertently exacerbate choice overload due to the nature of the choice sets they produce (Bollen et al., 2010; Diehl, 2005). If the resulting choice set consists of a large number of options that are all attractive and relevant to the consumer's needs, choice difficulty may arise (Bollen et al., 2010; Scheibehenne et al., 2010). Anyone who has spent an entire evening looking for a movie to watch on Netflix only to walk away defeated, finally settling for an old episode of the 90s TV show Friends, understands this.

Decision-making processes involving Artifical Intelligence-based RS can be seen as a collaboration between the system and the human, which calls for an understanding of how humans make choices and how these choices can be supported by algorithms (Gai & Klesse, 2019; Jameson et al., 2015; Puntoni et al., 2021). However, there has been a relative dearth of research studying the effect of RS algorithms on human behavior (Bollen et al., 2010; Mandl et al., 2011). Recommender systems researchers and designers have mostly focused on the technical side of their development, awarding less attention to the decision-making processes of the humans using them (Chen et al., 2013). Consumer psychology scholars have similarly neglected incorporating RS-produced choice sets into studies of consumer decision making, with most studies using standardized, rather than personalized, choice sets (Bollen et al., 2010).

The current research therefore aims to close this gap by studying consumer choice in a field setting involving a RS-produced choice set. Based on the theory and empirical evidence that show that reducing the attractiveness of the alternatives in a choice set facilitates choice (Bollen et al., 2010; Chernev et al., 2015; Scheibehenne et al., 2010; Willemsen et al., 2016), I manipulate the output of a RS algorithm accordingly. The revised algorithm output is more varied, such that consumers are presented with only the most attractive products followed by products of diminished appeal. In a randomized controlled field experiment consisting of 50,861 search sessions conducted by 23,165 consumers in two online retailers, no evidence is found for the occurrence of choice overload, or the mitigation of choice difficulty by the altered RS algorithm, when examining both process and outcome measures. Taken as a whole, this study proposes that studying consumer psychology phenomena in the unique setting provided by RS algorithms can result in both rich scientific insights and more effective algorithms.

Theoretical Background

Online retailers, unrestricted by the space confinements of brick and mortar retailers, offer seemingly endless assortments to their shoppers (J. Ma, 2016). To illustrate, a search on Amazon produced over 3,000 results for "toothpaste," over 10,000 results for "carpet," over 50,000 results for "dress," and over 100,000 results for "jacket."[1] The benefits of providing consumers with large choice sets are plentiful (Broniarczyk et al., 2018). To start, they are appealing to consumers, who much like economists, are attracted to choice (Loewenstein, 2000). Large assortments are perceived to contribute to a stimulating and pleasurable shopping experience, increase the likelihood of finding the ideal product for each individual consumer and provide an opportunity for variety-seeking behavior (Aurier & Mejía, 2020; Broniarczyk et al., 2018; Chernev et al., 2015). Indeed, when asked to choose between large and small assortments, consumers state a clear preference toward the former (Aydinli et al., 2017; Kahn & Lehmann, 1991; Natan, 2020; Oppewal & Koelemeijer, 2005). Overall, large assortments have been found to increase positive evaluation and preference, subsequently leading to increased purchase (Gao & Simonson, 2016; Sethuraman et al., 2022).

However, choosing from large assortments can also carry with it negative costs for consumers (Sethuraman et al., 2022). Large choice sets involve difficult trade-offs between alternatives and are therefore associated with higher cognitive load (Malhotra, 1982; Shugan, 1980). When choosing between an increased number of alternatives, consumers may revert to heuristic-based decision making (Payne et al., 1993), such as relying on fewer attributes in their evaluation (Timmermans, 1993) or choosing the most popular item (Sasaki et al., 2011). Consumers' level of certainty about their choice and commitment to it may be reduced (Dhar, 1997; Tversky & Shafir, 1992), and they are

[1] Search performed on September 17th, 2020. Amazon does not state the exact number of products that answer the search criteria, instead giving a rounded estimate as shown in these examples.

more likely to regret their choice (Inbar et al., 2011; Mannetti et al., 2007). The increased complexity associated with large choice sets may cause consumers to withdraw from what they perceive as an unpleasant decision process (Greenleaf & Lehmann, 1995; Iyengar & Lepper, 2000). This may therefore result in consumers deferring choice altogether (Dhar, 1996; Huang et al., 2018; Iyengar et al., 2004; Novemsky et al., 2007).

Choosing from large product assortments has therefore been associated with both greater benefits for consumers (sometimes referred to as more-is-better, e.g., Gao & Simonson, 2016) and higher costs (often called choice overload, see Schwartz & Ward, 2004). Which one of these effects is more likely to prevail in a choice situation involving large assortments has been the topic of a decades-long scholarly debate (Mochon & Schwartz, 2020). At least four meta-analyses have been dedicated to the concept of choice overload (Chernev et al., 2015; McShane & Böckenholt, 2018; Scheibehenne et al., 2010) or more generally to the effect of large assortments on consumer choice (Sethuraman et al., 2022). The first meta-analysis (Scheibehenne et al., 2010) found that overall, an increase in assortment size led to positive consequences (more-is-better). However, the authors conceded that choice overload may reliably occur in the presence of certain moderators related to the structure of the assortment (beyond the mere large quantity of options). The second meta-analysis (Chernev et al., 2015) went further to identify four specific moderators for the occurrence of choice overload in large assortments: choice set complexity, decision task difficulty, preference uncertainty, and decision goal. The third meta-analysis (McShane & Böckenholt, 2018) focused on improving the robustness of the methodology used by the previous two meta-analyses and came to largely similar conclusions as Chernev et al. (2015); namely, that choice overload exists but is likely to manifest only under certain boundary conditions. Finally, a recent meta-analysis on the effect of assortment size on consumer choice (Sethuraman et al., 2022) found evidence for a more-is-better effect. The authors concluded that overall larger assortments are beneficial to both consumers and retailers, while also acknowledging that this effect may be moderated by a host of environment- and study-related factors.

One contextual factor that seems to affect consumer response to product choice in large assortments is item attractiveness (Chernev et al., 2015; Scheibehenne et al., 2010). Choice sets containing highly attractive alternatives present a particular challenge to consumers (Summers, 1974; Wilkie, 1974). The cognitive load required to process a large choice set of highly attractive options is larger than when the attractiveness level is varied (Broniarczyk et al., 2018). Choice sets containing relatively equally attractive options therefore exacerbate consumer preference uncertainty and lead to greater choice deferral (Dhar, 1997), weakened choice accuracy (Malhotra, 1982), and decreased consumer satisfaction (Chan, 2015). Overall item attractiveness contributes to a high choice set complexity, which has been recognized as a driver for choice overload (Chernev et al., 2015).

The role of item attractiveness on determining choice difficulty suggests that choice sets produced by RS may be particularly relevant for the study of choice overload (Bollen et al., 2010). RS algorithms aim to provide consumers with highly attractive recommendations. Evaluating such options is cognitively effortful since it requires conducting trade-offs between appealing but incomparable features, which increases choice difficulty (Chernev et al., 2015; Scheibehenne et al., 2010). Therefore, choice sets produced by RS may be particularly susceptible to the occurrence of choice overload due to the high number of options coupled with their high attractiveness level. In a study that manipulated both the quantity and the quality of the alternatives in a RS-produced movie choice set, Bollen et al. (2010) found evidence to support this. The authors demonstrated that there is an increase in cognitive effort when consumers choose from a large choice set containing only attractive items compared with a large choice set containing items that vary in their attractiveness. The authors therefore proposed that incorporating qualitatively inferior options in the choice set may mitigate choice difficulty while not decreasing satisfaction, since the increased variety of the choice set increases the perception of its overall attractiveness. In follow-up work, Willemsen et al. (2016) showed that diversifying the items in a RS-produced choice set mitigated choice overload by reducing effort and choice difficulty, while maintaining overall attractiveness perception of the entire choice set.

Taken together, the surveyed literature signals that the personalized nature of the product lists produced by RS algorithms may result in a highly attractive and complex choice set (Chernev et al., 2015), making such choice sets a fertile ground for the occurrence of choice overload (Bollen et al., 2010; Willemsen et al., 2016). Given that online assortments offered by retailers are typically very large, varying the attractiveness level of items may be a promising avenue for decreasing choice difficulty and thereby choice overload. The current research aims to empirically test this idea by taking a similar approach to Bollen et al. (2010).[2] While keeping the overall size of the choice set constant, item attractiveness is varied such that only a small set of highly attractive choices is presented, followed by items of diminished appeal. The theoretical work discussed previously would predict that this intervention would mitigate choice overload, as measured by improvements to both choice outcome and process (Chernev et al., 2015; Reutskaja & Hogarth, 2009).

Method

This study was preregistered in Open Science Framework (https://osf.io/cm2rs). The preregistration documented the research question, variables, conditions, exclusion criteria, and analysis, and it was adhered to unless otherwise stated. The preregistration, data, code, and other materials are available in the Open Science Framework project (https://osf.io/vzfbw/).

Setting and Participants

Participants in the experiment were real-life shoppers in two e-commerce retailers. The experiment was initially conducted in a carpet retailer (henceforth Retailer A) and was later replicated in a photography equipment retailer (henceforth Retailer B). The retailers were chosen by the industry partner in this experiment, the creator of the RS algorithm used by these and other retailers. The retailers gave their support to conducting the experiment as per the researcher's plan, and their consent to sharing their data. The experiment ran for 4 weeks in each retailer during April–May, 2020. The length of the experiment was predetermined based on a power analysis suggesting that 1 month of traffic (approximately 50,000 search sessions) would be sufficient to detect a 5% increase in add-to-carts with 80% power and 5% significance level. A 5% conversion rate uplift in online shops is considered minimal to modest.

In total, 23,165 consumers performed 50,861 search sessions, resulting in 33,981 products viewed and 458 purchases. The retailer data were provided at the level of a search session. A search session is defined as the set of interactions, such as product clicks, add-to-carts and purchases, that a specific consumer undertakes after searching for a specific query. A total of 976 observations (search sessions) conducted by 204 consumers were excluded following the *2.5 SD above the mean* criteria recommended by Meyvis and Van Osselaer (2018), as preregistered. See overview of the sample in Table 1.

Procedure

The experiment had a between-subjects design. Under the duration of the study, consumers shopping at Retailers A and B who searched for a product or category using the search bar were randomly assigned to one of two conditions. Only consumers who proceeded with the default *rank by relevance*, were included in the experiment. Those who instead chose to rank the recommendation list by *price* or *news* were not included in the randomization process and their data were never collected, since they viewed nonpersonalized rankings. A thorough description of the algorithmic model used by the RS in this study can be found in Appendix A.

The two experimental conditions consumers were randomized into were *high Attractiveness (HA)* and *low Attractiveness (LA)*. Consumers in the *HA* condition received a search output in which products were ranked by the RS algorithm in decreasing order of fit based on relevance and personalization: Products most likely to appeal to the consumer were ranked highest, and products least likely to do so ranked lowest. Namely, the search output in this condition resembled the

[2] In their study, Bollen et al. (2010) were interested in differences in both quantity and quality of assortments, and their study included three conditions: a top-5 list (low quantity, high quality), a top-20 list (high quantity, high quality), and a Lin-20 list composed of the top-5 and an additional 15 lower ranked movies (high quantity, low quality). In the current case, since online assortments are almost always large, the focus was on differences in quality only, more specifically item attractiveness. The conditions of the present study correspond therefore to the two high quantity conditions in Bollen et al. (2010): top-20 and Lin-20.

Table 1
Overview of the Sample

Metric	Retailer A	Retailer B	Total
Consumers	7,928	15,237	23,165
Search sessions	14,805	36,056	50,861
Products purchased	273	185	458
Products added to cart	749	NA	749
Products viewed	14,538	19,443	33,981
Session length (s)	101.91	22.01	61.96

Note. Retailer B did not provide add-to-cart data. For session length, mean in seconds is shown instead of total. NA = not applicable.

output typically produced by RS.[3] Consumers in the *LA* condition received an altered search output. Given that previous research has established an association between high item attractiveness and choice difficulty (Bollen et al., 2010; Chernev et al., 2015; Dhar & Wertenbroch, 2000; Malhotra, 1982), this condition aimed to reduce item attractiveness in order to mitigate choice overload. The search output in this condition featured the products predicted to have the highest consumer fit at the top seven spots, but it followed by displaying the products predicted by the algorithm to have the *lowest* fit, taken from the bottom of the ranking list. For consumers randomized into this condition, those poorly fit products were now featured prominently, in places eight onwards. As argued by Bollen et al. (2010), this lower quality condition should result in reduced trade-off difficulty, as there would be less "competition" between the attractive top items, and the less appealing items that follow.[4]

Keeping highly attractive items in the top spots even in the *LA* condition was done in order to avoid customer confusion and frustration. Consumers typically examine the top results first, regardless of their quality (Jameson et al., 2015; Ursu, 2018). Having unattractive products in the top spots would therefore risk baffling consumers and be unrealistic in this field setting. Therefore, the product ranking produced by the algorithm in the *LA* condition contained attractive products as the top seven results, representing around 30% of the product results displayed on the first search output page (24). This rate was chosen to balance maintaining consumer satisfaction along with allowing for sufficient treatment intensity to test the prediction that attractiveness would affect choice overload.

It should be noted that while the products displayed in places eight onwards in the LA condition were taken from the bottom of the product list produced by the algorithm, they were still *relevant* products. In other words, if a customer in Retailer B searched for *tripod*, the results at the bottom of the ranking were less common versions of *tripods*, rather than, for example, *binoculars*, *drones,* or *film-rolls* (all products categories sold by Retailer B). Figure 1 shows an example of a search output for the same term in both HA and LA conditions.

Outcome Measures

Choice overload was measured by looking at both choice outcome and process, following, for example, Bollen et al. (2010); Chernev et al. (2015); Reutskaja and Hogarth (2009). To judge whether a successful outcome was achieved, I looked at whether a decision was made, by measuring whether the session led to a product purchased (Model 1) or at least added to cart (Model 2). To judge whether the process had been effective, I looked at number of products viewed (Model 3), conjecturing that a larger number indicates a higher level of choice overload due to the need to evaluate many options (as was assumed by Bollen et al., 2010). In assessing the process, I also looked at session time (Model 4), assuming that increased time points to increased choice overload due to the effort required to make a decision.

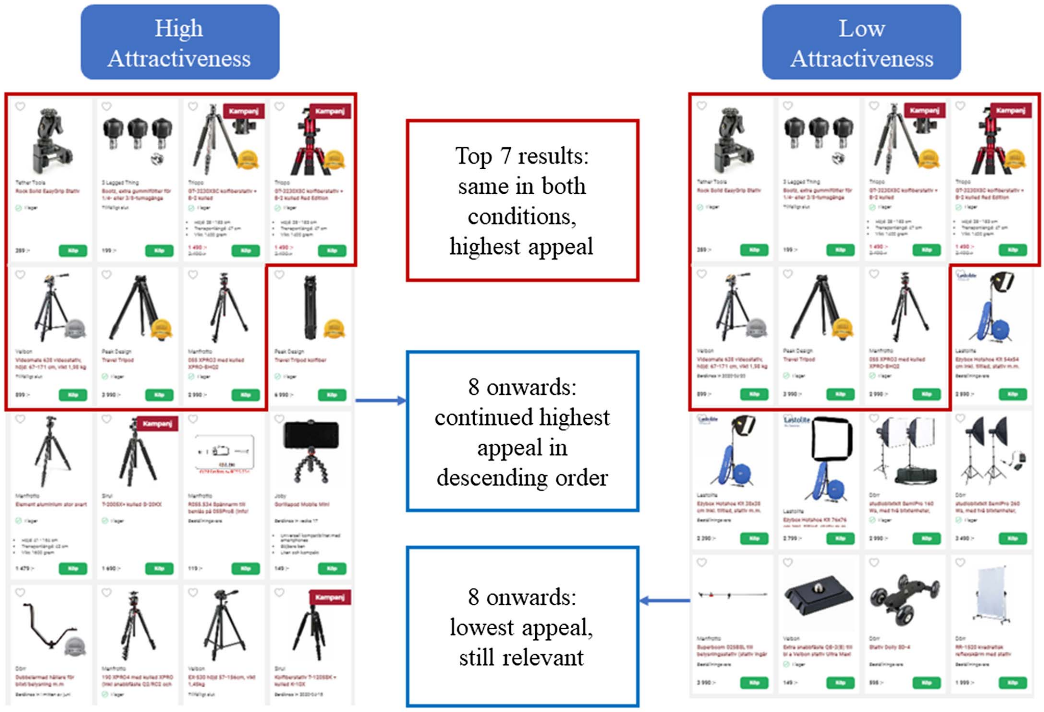
Analysis

To model consumer choices based on the data different variations of linear mixed-effects regression models were employed (Gill & Womack, 2013). Mixed-effects models, also known as multilevel or hierarchical models, are commonly used on nested data, where observations are in hierarchical structures (Gelman & Hill, 2006). In this case, the data were nested in two levels:

[3] The *high attractiveness (HA)* condition represents the typical RS algorithm used by the industry partner in other online stores. Experiments conducted by the industry partner to test the efficacy of this algorithm have found it to increase conversion rate (percentage of visitors whose search resulted in a purchase) by an average of 8.2% versus a nonpersonalized algorithm.

[4] This condition is inspired by the Lin-20 condition of Bollen et al. (2010) which was composed of the top-5 and an additional 15 lower ranked movies and represented a high quantity, lower quality condition.

Figure 1
Search Output for Term "Tripod" ("Stativ") in Both Conditions



Note. See the online article for the color version of this figure.

The observation level was a *search session* while the randomization occurred on the level of the individual *consumer*. One consumer could perform several search sessions, the number of which varied from individual to individual. This nested structure of the data implies the lack of independence between search sessions belonging to the same consumer, which allowing for random effects on the consumer level accounts for. Therefore, multilevel models were employed by adding a random intercept by consumer, whereby search sessions were specified at Level 1 and consumers were specified at Level 2.

General Model Specification

Four different multilevel regression models were applied as determined by each outcome measure in Outcome Measures section. To illustrate the general model, I therefore focus on the right-hand side containing the predictors, which was common for all models. Particularities stemming from the outcome variable for each model will be outlined in the following Specific Models section. Note that as random assignment was employed, no control variables were added to the models. The two-level model with random intercepts was estimated as follows:

$$\text{Level 1}: y_{ij} = \beta_{0j} + \beta_1 \times \text{Condition}_j + \varepsilon_{ij}, \quad (1)$$

where y_{ij} represented the specific's model outcome variable for the *i*th search session from the *j*th consumer, $Condition_j$ was the treatment variable with the values *HA* and *LA* and behavior intrinsic to an individual consumer β_{0j} was modeled as a random effect model via this equation:

$$\text{Level 2}: \beta_{0j} = \beta_{00} + U_{0j}, \quad (2)$$

where β_{00} was a fixed effect and U_{0j} accounted for (random) within-consumer variation by search session. On an aggregate level, the model can be illustrated as:

$$y_{ij} = \beta_{00} + \beta_1 \times \text{Condition}_j + U_{0j} + \varepsilon_{ij}, \quad (3)$$

where y_{ij} represented the specific's model outcome variable for the ith search session from the jth consumer, Condition$_j$ was the treatment, U_j the random intercept for the jth consumer at Level 2, and ε_{ij} was the overall error term.

Specific Models

Model 1 was a multilevel logistic regression model estimated using maximum likelihood (ML) approach where outcome variable y_{ij}^{model1} was a binary variable for the occurrence of purchase. The variable was coded as 1 = occurrence of purchase, 0 = no occurrence of purchase. Model 2 was a multilevel logistic regression model estimated using ML approach where outcome variable y_{ij}^{model2} was a binary variable for the occurrence of add-to-cart. The variable was coded as 1 = occurrence of add-to-cart, 0 = no occurrence of add-to-cart. Model 3 was a multilevel negative binomial regression estimated using ML estimation approach where outcome variable y_{ij}^{model3} was the number of products viewed. Model 4 was a multilevel linear regression estimated using restricted maximum likelihood estimation approach where outcome variable y_{ij}^{model4} was session time in seconds.

As a rule, generalized linear mixed models were estimated using ML estimation approach, whereas linear mixed models were estimated using restricted maximum likelihood estimation approach, as is generally preferred (Gilmour et al., 1995; McCulloch, 1997). Applying multilevel models represented a deviation from the preregistered analysis plan which did not account for the dependencies in the data, as data structure was not known at the time of preregistration.

Results

Choice overload was operationalized as a lower likelihood of occurrence of purchase/add-to-cart, longer time-to-decision, and a more laborious decision process, as described in Outcome Measures section. A summary of descriptive statistics by condition is presented in Table 2. For all models, fixed effects of the manipulation are reported. Odd ratios (*OR*) and incidence rate ratios (IRR) were computed with the 95% confidence interval in brackets. All analyses reported were conducted via the statistical software R (v.4.0.3). Full results are presented in Table 3.

Table 2

Descriptive Data by Condition

Outcome variable	*M*		*SD*		Min		Max		N observation		N consumer	
	HA	LA	HA	LA	HA	LA	HA	LA	HA	LA	HA	LA
Retailer A												
Number of purchases	0.0185	0.0184	0.143	0.15	0	0	2	3	7,305	7,500	3,949	3,979
Number of add-to-carts	0.0517	0.0495	0.254	0.263	0	0	6	6	7,305	7,500	3,949	3,979
Number of products viewed	0.986	0.978	1.7	1.82	0	0	34	50	7,305	7,500	3,949	3,979
Session length	102.943	100.897	243.281	231.436	0	0	4,701	4,325	7,305	7,500	3,949	3,979
Retailer B												
Number of purchases	0.00482	0.00545	0.0708	0.0766	0	0	2	3	18,264	17,792	7,748	7,489
Number of add-to-carts	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number of products viewed	0.534	0.544	0.839	0.849	0	0	15	16	18,264	17,792	7,748	7,489
Session length	22.214	21.804	56.038	55.552	0	0	408	418	18,264	17,792	7,748	7,489

Note. Retailer B did not provide add-to-cart data. Session length is shown in seconds. Descriptive data given on a search session level and does not account for dependencies in the data. N observations means the total number of observations and N consumers means the total number of consumers. HA = high attractiveness; LA = low attractiveness; Min = minimum; Max = maximum; NA = not applicable.

Table 3
Multilevel Models for Decision Outcome and Process

Effect	*B*	Statistic	*p*	Exp (*B*)	95% CI *LL*	95% CI *UL*
Model 1: Purchase						
Retailer A						
Condition (LA)	0.14	0.44	.662	1.15	−0.49	0.78
Intercept	−11.25	−5357.05	<.001	0	−11.3	−11.2
Retailer B						
Condition (LA)	0.23	0.58	.559	1.26	−0.55	1.01
Intercept	−13	−6477.2	<.001	0	−13	−13
Model 2: Add-to-cart						
Retailer A						
Condition (LA)	0	0.003	.998	1	−0.55	0.56
Intercept	−9.63	−35.34	<.001	0	−10.2	−9.1
Retailer B						
Condition (LA)	NA	NA	NA	NA	NA	NA
Intercept	NA	NA	NA	NA	NA	NA
Model 3: Products viewed						
Retailer A						
Condition (LA)	0.014	0.46	.645	1.01	−0.04	0.07
Intercept	−0.35	−15.14	<.001	0.7	−0.4	−0.3
Retailer B						
Condition (LA)	0.014	0.75	.454	1.01	−0.02	0.05
Intercept	0.47	−53.04	<.001	0.48	−0.76	−0.7
Model 4: Session length						
Retailer A						
Condition (LA)	2.06	0.47	.635		−6.46	10.58
Intercept	101.64	32.95	<.001		95.6	107.69
Retailer B						
Condition (LA)	−0.46	−0.74	.46		−1.69	0.76
Intercept	22.47	51.15	<.001		21.6	23.3

Note. Odds ratio (*OR*) is reported for logistic regression (Models 1 and 2). Incidence rate ratio is reported for negative binomial regression (Model 3). T-statistic is reported for linear regression (Model 4). Retailer B did not provide add-to-cart data. Exp (*B*) = exponential value of B; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; LA = low attractiveness; NA = not applicable.

Effect on Choice Outcome

To examine whether the intervention affected choice outcome, the effect of condition (HA/LA) on the likelihood of occurrence of a purchase (Model 1, see Table 3) and add-to-cart (Model 2, see Table 3) was tested. To note, only Retailer A shared add-to-cart data, therefore, there is no analysis for add-to-cart data for Retailer B.

In Retailer A, there was no significant difference between conditions for purchase ($B = 0.14$, $OR = 1.15$, 95% CI [−0.49, 0.78], $z = 0.44$, $p = .662$). There was likewise no significant difference between conditions for add-to-cart ($B = 0.00$, $OR = 1.00$, 95% CI [−0.55, 0.56], $z = 0.003$, $p = .998$). Similarly, in Retailer B, there was no significant difference between conditions for purchase ($B = 0.23$, $OR = 1.26$, 95% CI [−0.55, 1.016], $z = 0.58$, $p = .559$). Retailer B did not provide add-to-cart data.

Effect on Choice Process

To examine the effect of the intervention on choice process, I looked at whether the revised algorithm resulted in a less cognitively demanding (as evident by reduced need to "click around") and less time-consuming (as evident by a quicker path to decision) process. Similar measures were used by Bollen et al. (2010) to gauge choice difficulty.

Model 3 therefore tested the effect of condition (HA/LA) on the number of products viewed (see Table 3). In Retailer A, there was no significant difference between conditions ($B = 0.014$, IRR = 1.014, 95% CI [−0.04, 0.07], $z = 0.46$, $p = .645$). There was likewise no significant difference between conditions in Retailer B ($B = 0.014$, IRR = 1.014, 95% CI [−0.02, 0.05], $z = 0.75$, $p = .454$).

Model 4 tested the effect of condition (HA/LA) on the search session time (see Table 3). Also here, there was no significant difference between conditions in Retailer A, $t(14, 801) = 0.474$, $SE = 4.35$, $p = .635$. There was likewise no significant difference between conditions in Retailer B, $t(36, 052) = -0.74$, $SE = 0.63$, $p = .46$.[5]

Discussion

General Discussion

Recommender systems aim to assist consumers in choosing products from large assortments. However, the personalized nature of their output may result in choice sets with high item attractiveness, inadvertently engendering choice overload (Bollen et al., 2010). There is therefore a growing need for additional empirical research that examines choice overload in the specific context of RS. The present study aims to do just that, by examining whether changes to the design of a RS algorithm to reduce item attractiveness would affect choice process and outcome. The revised algorithm presented a search output with varied item attractiveness so that only few highly attractive options were presented at the top, followed by not-so-attractive ones. Testing this intervention in a real-life online shopping setting in two separate retailers, no significant differences were found in either decision outcome (as measured by product purchases and add-to-carts) or decision process (as measured by number of products viewed and session time).

To understand the study's null findings better under the frequentist null hypothesis significance testing framework, I followed up the results with conducting equivalence tests (Lakens et al., 2018)[6] for Models 1–4 previously summarized in Table 3 in Results section. The results were such that for choice process measures (product views, session length) equivalence was "accepted," while for choice outcome measures (purchases, add-to-carts) equivalence was "undecided." These results indicate that for the choice process measures, the confidence interval of the estimate were completely inside the region of practical significance (*ROPE*), which can be interpreted as having "practically no effect," or an effect of a negligible magnitude. However, for the choice outcome measures, the result were inconclusive and it is not possible to neither accept nor reject equivalence. In other words, the equivalence tests provide support for accepting the null for the choice process measures, though not for the choice outcome measures. See elaboration and full results of equivalence tests in Appendix B.

There are several additional factors that contribute to the study's robustness and may serve to increase confidence in the current null findings. First, the experiment had a large sample size (total 50,861 observations from 23,165 consumers), which was predetermined based on an ex ante power analysis as reported in the Setting and Participants section under Method. This addresses the concern of observing a null due to insufficient power to detect an effect. Second, the experiment used several methods of triangulation, recommended by Cortina and Folger (1998) as a basis for establishing that a good effort has been made to accept the null; the current experiment included different operationalizations of the outcome variables, and it was conducted in two separate settings using two independent samples. Finally, it would be reasonable to wonder if the null effect was a result of a failed manipulation. Perhaps consumers only focused on the first seven products to make their choice, never noticing the high/low level of attractiveness of the rest of the product list. Since these top products were highly attractive in both conditions, if consumers fixated only on them, it would effectively mean the manipulation was not noticeable. However, the data show that this is not the case. In Retailer A, out of a total of 14,538 products viewed (clicked on), 8,132 (56%) did not appear in the top seven results. Out of a total of 273 products purchased, 69 (25%) did not appear in the top seven results.[7] These numbers demonstrate that consumers did look beyond the top seven

[5] For the variable of search session time, an additional analysis was conducted with a log transformed variation, as helpfully suggested by reviewers. The results were likewise non statistically significant in neither Retailer A, $t(14,801) = 0.40$, $SE = 0.04$, $p = .686$, nor Retailer B, $t(36,052) = 0.80$, $SE = 0.02$, $p = .423$.

[6] I thank the reviewer who suggested to conduct this additional analysis.

[7] Retailer B did not provide product positions for total clicks and purchases.

results, both when considering which products to buy and when doing the actual purchasing, indicating that the manipulation was in fact noticeable. There were no significant differences between conditions in the distribution of product views and purchases from top (1–7)/bottom (8 and onwards) items. The full analysis is included in Appendix C.

Taken together, I argue that the experiment's high statistical power, its outcome variable and sample triangulation, and the evidence for the effectiveness of its manipulation assuage potential doubts that measurement error created the null effect, particularly for the choice process measures where equivalence was accepted.

Theoretical Contribution

This study adds to the literature in several ways. First, machine learning-based RS algorithms have become ubiquitous in online shopping, present in both retail giants such as Amazon and Walmart, and lesser known, small mom-and-pops brand stores. These recommendation algorithms have been shown to affect online consumer behavior in copious ways and have therefore attracted much research attention (Zhang et al., 2022). The current research adds to this body of knowledge by investigating whether reducing item attractiveness of an RS-produced product set would affect consumer choice. Prior research has primarily examined the effect of RS on choice overload in experiments conducted in lab settings (albeit realistic, e.g., Bollen et al., 2010; Willemsen et al., 2011, 2016). The current work is, to the author's knowledge, the first to examine the effect of recommender systems on choice overload in a field experiment conducted with consumers making real-life product choices in online stores. Importantly, the participants in the study were not aware that they were being observed, a key feature of a field experiment that affords a high degree of ecological value (Gneezy, 2017; van Heerde et al., 2021) which strengthens its relevance and contribution. In general, research in choice overload and assortment size has tended to happen in controlled lab environments, with fewer than 10% of empirical studies conducted in the field, which calls for diversifying our research paradigms and settings to reflect consumers' reality (Sethuraman et al., 2022).

Second, the study's findings may be interpreted to suggest that lowering item attractiveness did not facilitate choice as hypothesized. Further, the results may indicate that even a reduction as large as 70% in a choice set's item attractiveness did not hurt choice process or outcome. The revised algorithm neither decreased nor increased choice overload, despite the considerable mitigation of the attractiveness of the choice set's items. These findings may suggest that a small set of great options followed by many more not-so-great options could be as effective as a large set of only great options. These findings provide relevant insights for future research looking to disentangle the specific role of choice set attractiveness on choice overload. Previous research has suggested that reducing item attractiveness increased perceived variety without hindering choice (Bollen et al., 2010). While the current work has not found evidence that reduced item attractiveness *improved* choice outcome or process, it has likewise not found evidence supporting the prevailing conventional wisdom that attractive choice sets perform better (Chernev & Hamilton, 2009). An interesting direction to explore in a field setting would be increasing the *diversity* of the items in a choice set rather than reducing their *attractiveness*. Previous work from lab experiments in the film domain suggests that increasing item diversity reduces choice difficulty, while maintaining the perceived attractiveness of the choice set and consumers' satisfaction with it (Willemsen et al., 2016).

Finally, this work makes a methodological contribution to the study of consumer choice online. While most consumer choice research so far has been conducted with fixed, standardized choice sets, this work proposes that to effectively reflect the conditions of real-life consumer choice "in the wild," researchers should dare to use dynamic, personalized choice sets produced by RS algorithms. These sets are not only more reflective of choice sets consumers meet in their regular online shopping trips, but they also afford the researcher the ability to control features of product assortment in a systematic way (Bollen et al., 2010; Willemsen et al., 2016). In the present study, the RS setting allowed to control and limit the RS output's item attractiveness, enabling an investigation of the role of choice set complexity, identified by Chernev et al. (2015) as an important moderating factor of choice overload. This shows that RS algorithms can be a valuable asset for investigating phenomena such as choice overload, and more broadly, consumer behavior and decision making in shopping situations occurring online.

Practical Implications

As opposed to bricks and mortar stores, online retailers cannot employ human sales experts to support their shoppers in the buying process. Instead, many retailers implement some form of RS to help guide consumers toward the right products for them—those most likely to fit their needs and desires—in the hope that these algorithms would help facilitate choice (Marchand & Marx, 2020). However, the present work paints a more nuanced picture regarding RS algorithms. The findings concerning the role of attractiveness of the recommended products on customers' decision process and outcome present novel insights which should be considered when designing an effective RS for online retailers. Focusing on the attractiveness of the top choices seems to be equally effective to maximizing the attractiveness of the entire product set. This can provide guidance for both retailers and RS algorithm designers on where to direct their focus.

Further, this work demonstrates that incorporating psychological insights in the design of RS algorithms could be useful, as others have previously argued (Bollen et al., 2010; Chen et al., 2013; Jameson et al., 2015; L. Ma & Sun, 2020; Mandl et al., 2011; Puntoni et al., 2021; Willemsen et al., 2016) Considering human aspects of decision making and consumer behavior could help bridge the gap between RS algorithms' purpose to facilitate choice, and the reality of potential choice difficulty. The present study challenges the assumption held by many retailers that RS algorithms must create long lists of ranked products, as the results seem to suggest that a small set of highly attractive alternatives followed by many less attractive options can be as effective. Online retailers using RS may want to focus on the attractiveness of the top choices in a product set, rather than pay the same level of attention for the attractiveness level of the list as a whole.

Limitations and Future Research Directions

Although this research advances our understanding of RS and consumer choice, some limitations should be noted. First, this work examined the effect of RS on choice overload in two shopping categories: carpets and camera equipment. Previous research has suggested that the existence of choice overload can vary by product category, even when there is a lack of an overall effect (Ong, 2021; Sethuraman et al., 2022). Therefore, expanding the investigation to additional categories seems wise. Moreover, both categories used in this research can be considered utilitarian product categories (Babin & Darden, 1995; Khan et al., 2005). Yet, consumer online behavior may be different in products that are categorically different, for example, more hedonistic products, such as fashion. Future research should therefore explore additional choice environments where RS algorithms are involved in decision making, both within-consumer products (fashion, groceries) and beyond (movies, travel). Specifically, exploring multiple product categories that differ in hedonic/utilitarian attribute trade-offs might lead to valuable insights due to evidence that this trade-off affects attitudes toward Artifical Intelligenc-based recommendations (Longoni & Cian, 2022).

Second, the field nature of this study allowed to measure choice overload by looking at behavioral data: clicks, purchases, time spent. While this is a clear strength of the study in terms of ecological validity, this field setting did not allow to capture other aspects which would be of interest, for example, consumer attitudes such as satisfaction or regret. These attitudes, and the consumer's internal state more broadly, are important manifestations of choice overload (Chernev et al., 2015; Schwartz & Ward, 2004), but were not accessible in this study's particular setting. Future research should strive to combine subjective and objective measures of choice overload, for example, by combining a field experiment conducted in an online shop with postpurchase survey questions.

Finally, due to the inconclusive nature of the null findings of this work, a call for replication in order, with an even larger sample to increase power. Ideally, such a replication can also extend the work by exploring potentially interesting heterogeneous effects (Bryan et al., 2021) There is reason to believe that consumer behavior online would be affected by several individual inclinations such as the consumer's level of product category engagement or expertise. Future research could investigate this in a similar yet more powered field study, for example, by looking at different shoppers' purchase history, or number of previous visits in the same online store.

Conclusion

This field experiment with 23,165 customers in two online shopping platforms tested the effect of reducing item attractiveness of a choice set produced by a machine learning, RS algorithm, on choice outcome and process. The revised algorithm presented only the top highly attractive results, followed by results of lower appeal. The intervention did not result in any statistically significant changes to either decision outcome, as measured by products purchased or added to cart, or decision process, as measured by number of products viewed, and search session length. These results, and additional equivalence analysis, suggest that this considerable reduction in a choice set's attractiveness did not hurt consumer choice, indicating that a small set of highly attractive products followed by many less appealing options could work as effectively as a large set of only attractive options. These findings are important in light of the growing number of decisions consumers make online, and the pivotal role RS algorithms increasingly play in these decisions. This research therefore demonstrates that incorporating psychological insights in the design of RS algorithms could be beneficial for both theory and practice.

References

Aurier, P., & Mejía, V. D. (2020). The interplay of brand-line assortment size and alignability in the sales of brand-lines and line-extensions of frequently purchased products. *Journal of Business Research*, *117*, 163–175. https://doi.org/10.1016/j.jbusres.2020.05.024

Aydinli, A., Gu, Y., & Pham, M. T. (2017). An experience-utility explanation of the preference for larger assortments. *International Journal of Research in Marketing*, *34*(3), 746–760. https://doi.org/10.1016/j.ijresmar.2017.06.007

Babin, B. J., & Darden, W. R. (1995). Consumer self-regulation in a retail environment. *Journal of Retailing*, *1*(71), 47–70. https://doi.org/10.1016/0022-4359(95)90012-8

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., & Graus, M. (2010). *Understanding choice overload in recommender systems* [Conference session]. Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 10, New York, NY, United States. Association for Computing Machinery.

Broniarczyk, S. M., Chernev, A., Duhachek, A., Goodman, J., & Griffin, J. (2018). Product assortment. In C. P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology* (pp. 755–779). Routledge.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Burges, C. J. (2010). From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, *11*(23–581), 81.

Chan, E. Y. (2015). Attractiveness of options moderates the effect of choice overload. *International Journal of Research in Marketing*, *32*(4), 425–427. https://doi.org/10.1016/j.ijresmar.2015.04.001

Chen, L., de Gemmis, M., Felfernig, A., Lops, P., Ricci, F., & Semeraro, G. (2013). Human decision making and recommender systems. *ACM Transactions on Interactive Intelligent Systems*, *3*(3), Article 17. https://doi.org/10.1145/2533670.2533675

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, *25*(2), 333–358. https://doi.org/10.1016/j.jcps.2014.08.002

Chernev, A., & Hamilton, R. (2009). Assortment size and option attractiveness in consumer choice among retailers. *Journal of Marketing Research*, *46*(3), 410–420. https://doi.org/10.1509/jmkr.46.3.410

Cortina, J. M., & Folger, R. G. (1998). When is it acceptable to accept a null hypothesis: No way, Jose? *Organizational Research Methods*, *1*(3), 334–350. https://doi.org/10.1177/109442819813004

Dhar, R. (1996). The effect of decision strategy on deciding to defer choice. *Journal of Behavioral Decision Making*, *9*(4), 265–281. https://doi.org/10.1002/(SICI)1099-0771(199612)9:4<265::AID-BDM231>3.0.CO;2-4

Dhar, R. (1997). Consumer preference for a no-choice option. *Journal of Consumer Research*, *24*(2), 215–231. https://doi.org/10.1086/209506

Dhar, R., & Wertenbroch, K. (2000). Consumer choice between hedonic and utilitarian goods. *Journal of Marketing Research*, *37*(1), 60–71. https://doi.org/10.1509/jmkr.37.1.60.18718

Diehl, K. (2005). When two rights make a wrong: Searching too much in ordered environments. *Journal of Marketing Research*, *42*(3), 313–322. https://doi.org/10.1509/jmkr.2005.42.3.313

Gai, P. J., & Klesse, A.-K. (2019). Making recommendations more effective through framings: Impacts of user- versus item-based framings on recommendation click-throughs. *Journal of Marketing*, *83*(6), 61–75. https://doi.org/10.1177/0022242919873901

Gao, L., & Simonson, I. (2016). The positive effect of assortment size on purchase likelihood: The moderating influence of decision order. *Journal of Consumer Psychology*, *26*(4), 542–549. https://doi.org/10.1016/j.jcps.2015.12.002

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gill, J., & Womack, A. J. (2013). The multilevel model framework. In M. A. Scott, B. D. Marx, &

J. S. Simonoff (Eds.), *The SAGE handbook of multilevel modeling* (pp. 3–20). SAGE Publications.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, *51*, 1440–1450. https://doi.org/10.2307/2533274

Gneezy, A. (2017). Field experimentation in marketing research. *Journal of Marketing Research*, *54*(1), 140–143. https://doi.org/10.1509/jmr.16.0225

Greenleaf, E. A., & Lehmann, D. R. (1995). Reasons for substantial delay in consumer decision making. *Journal of Consumer Research*, *22*(2), 186–199. https://doi.org/10.1086/209444

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, *19*(1), 4–21. https://doi.org/10.1287/mksc.19.1.4.15178

He, X., Zhang, H., Kan, M.-Y., & Chua, T.-S. (2016). *Fast matrix factorization for online recommendation with implicit feedback* [Conference session]. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, United States. https://doi.org/10.1145/2911451.2911489

Huang, G.-H., Korfiatis, N., & Chang, C.-T. (2018). Mobile shopping cart abandonment: The roles of conflicts, ambivalence, and hesitation. *Journal of Business Research*, *85*, 165–174. https://doi.org/10.1016/j.jbusres.2017.12.008

Inbar, Y., Botti, S., & Hanko, K. (2011). Decision speed and choice regret: When haste feels like waste. *Journal of Experimental Social Psychology*, *47*(3), 533–540. https://doi.org/10.1016/j.jesp.2011.01.011

Iyengar, S. S., Huberman, G., & Jiang, W. (2004). How much choice is too much? Contributions to 401(k) retirement plans. In O. S. Mitchell & S. P. Utkus (Eds.), *Pension design and structure: New lessons from behavioral finance* (pp. 83–97). Oxford University Press.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

Jameson, A., Willemsen, M. C., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., & Chen, L. (2015). Human decision making and recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 611–648). Springer.

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems: An introduction*. Cambridge University Press.

Kahn, B. E., & Lehmann, D. R. (1991). Modeling choice among assortments. *Journal of Retailing*, *67*(3), 274–299.

Khan, U., Dhar, R., & Wertenbroch, K. (2005). A behavioral decision theory perspective on hedonic and utilitarian choice. In S. Ratneshwar, & D. G. Mick (Eds.), *Inside consumption: Frontiers of research on consumer motives, goals, and desires* (Vol. 1, pp. 144–165). Routledge.

Köcher, S., Jugovac, M., Jannach, D., & Holzmüller, H. H. (2019). New hidden persuaders: An investigation of attribute-level anchoring effects of product recommendations. *Journal of Retailing*, *95*(1), 24–41. https://doi.org/10.1016/j.jretai.2018.10.004

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta- analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Loewenstein, G. (2000). Costs and benefits of health- and retirement-related choice. In S. Burke, E. Kingson, & U. Reinhardt (Eds.), *Social security and medicare: Individual vs. collective risk and responsibility* (pp. 87–100). Brookings Institution Press.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Ma, J. (2016). Does greater online assortment pay? An empirical study using matched online and catalog shoppers. *Journal of Retailing*, *92*(3), 373–382. https://doi.org/10.1016/j.jretai.2016.05.004

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), Article 1541. https://doi.org/10.21105/joss.01541

Malhotra, N. K. (1982). Information load and consumer decision making. *Journal of Consumer Research*, *8*(4), 419–430. https://doi.org/10.1086/208882

Mandl, M., Felfernig, A., Teppan, E., & Schubert, M. (2011). Consumer decision making in knowledge-based recommendation. *Journal of Intelligent Information Systems*, *37*(1), 1–22. https://doi.org/10.1007/s10844-010-0134-3

Mannetti, L., Pierro, A., & Kruglanski, A. (2007). Who regrets more after choosing a non-status-quo option?

Post decisional regret under need for cognitive closure. *Journal of Economic Psychology*, *28*(2), 186–196. https://doi.org/10.1016/j.joep.2005.12.004

Marchand, A., & Marx, P. (2020). Automated product recommendations with preference-based explanations. *Journal of Retailing*, *96*(3), 328–343. https://doi.org/10.1016/j.jretai.2020.01.001

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, *92*(437), 162–170. https://doi.org/10.1080/01621459.1997.10473613

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1), 255–271. https://doi.org/10.1007/s11336-017-9571-z

Meyvis, T., & Van Osselaer, S. M. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research*, *44*(5), 1157–1173. https://doi.org/10.1093/jcr/ucx110

Mochon, D., & Schwartz, J. (2020). The importance of construct validity in consumer research. *Journal of Consumer Psychology*, *30*(1), 208–214. https://doi.org/10.1002/jcpy.1145

Natan, O. R. (2020). *Choice frictions in large assortments*. University of Chicago.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research*, *44*(3), 347–356. https://doi.org/10.1509/jmkr.44.3.347

Ong, D. (2021). Predicting choice-averse and choice-loving behaviors in a field experiment with actual shoppers. *Journal of Economic Behavior & Organization*, *188*, 46–71. https://doi.org/10.1016/j.jebo.2021.05.002

Oppewal, H., & Koelemeijer, K. (2005). More choice is better: Effects of assortment size and composition on assortment evaluation. *International Journal of Research in Marketing*, *22*(1), 45–60. https://doi.org/10.1016/j.ijresmar.2004.03.002

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

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Reutskaja, E., & Hogarth, R. M. (2009). Satisfaction in choice as a function of the number of alternatives: When "goods satiate". *Psychology & Marketing*, *26*(3), 197–203. https://doi.org/10.1002/mar.20268

Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. Kantor (Eds.), *Recommender systems handbook* (pp. 1–35). Springer.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Sasaki, T., Becker, D. V., Janssen, M. A., & Neel, R. (2011). Does greater product information actually inform consumer decisions? The relationship between product information quantity and diversity of consumer decisions. *Journal of Economic Psychology*, *32*(3), 391–398. https://doi.org/10.1016/j.joep.2011.02.010

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, *37*(3), 409–425. https://doi.org/10.1086/651235

Schwartz, B., & Ward, A. (2004). Doing better but feeling worse: The paradox of choice. In P. A. Linley & S. Joseph (Eds.), *Positive psychology in practice* (pp. 86–104). Wiley.

Sethuraman, R., Gázquez-Abad, J. C., & Martínez-López, F. J. (2022). The effect of retail assortment size on perceptions, choice, and sales: Review and research directions. *Journal of Retailing*, *98*, 24–45. https://doi.org/10.1016/j.jretai.2022.01.001

Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research*, *7*(2), 99–111. https://doi.org/10.1086/208799

Summers, J. O. (1974). Less information is better? *Journal of Marketing Research*, *11*(4), 467–468. https://doi.org/10.1177/002224377401100415

Swaminathan, A., & Joachims, T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, *16*(1), 1731–1755.

Timmermans, D. (1993). The impact of task complexity on information use in multi-attribute decision making. *Journal of Behavioral Decision Making*, *6*(2), 95–111. https://doi.org/10.1002/bdm.3960060203

Tversky, A., & Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, *3*(6), 358–361. https://doi.org/10.1111/j.1467-9280.1992.tb00047.x

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, *37*(4), 530–552. https://doi.org/10.1287/mksc.2017.1072

van Heerde, H. J., Moorman, C., Moreau, C. P., & Palmatier, R. W. (2021). Reality check: Infusing ecological value into academic marketing research. *Journal of Marketing*, *85*(2), 1–13. https://doi.org/10.1177/0022242921992383

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., & Najork, M. (2018). *Position bias estimation for unbiased learning to rank in personal search* [Conference session]. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, New York, NY, United States. https://doi.org/10.1145/3159652.3159732

Wilkie, W. L. (1974). Analysis of effects of information load. *Journal of Marketing Research*, *11*(4), 462–466. https://doi.org/10.1177/002224377401100414

Willemsen, M. C., Graus, M. P., & Knijnenburg, B. P. (2016). Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, *26*(4), 347–389. https://doi.org/10.1007/s11257-016-9178-6

Willemsen, M. C., Knijnenburg, B. P., Graus, M. P., Velter-Bremmers, L., & Fu, K. (2011). Using latent features diversification to reduce choice difficulty in recommendation lists. In A. Felfernig, L. Chen, M. Mandl, M. Willemsen, D. Bollen, & M. Ekstrand (Eds.), *The RecSys 2011 Workshops–Decisions@RecSys 2011 and UCERSTI-2* (pp. 14–20). CEUR-WS.org.

Zhang, J., Balaji, M., Luo, J., & Jha, S. (2022). Effectiveness of product recommendation framing on online retail platforms. *Journal of Business Research*, *153*, 185–197. https://doi.org/10.1016/j.jbusres.2022.08.006

Appendix A

The Algorithmic Model

This section provides a more detailed description of the algorithmic model used in the study. The model includes two subsequent stages: The first, *the informational retrieval system*, includes the retrieval and ranking of relevant items based on a customer's search query and is also called *machine-learned ranking*. The second, *the recommender system*, includes the recommendation of the retrieved products based on user and item attributes. In the first stage, the algorithm assigns a *relevancy* score to each product based on its comparability with a given consumer search query ("How similar is this product to *tripod*?"). In the second stage, the algorithm relies on both explicit (e.g., product ratings) and implicit (e.g., browsing patterns) customer input to produce a *personalized* product ranking ("How likely is this particular tripod to appeal to this specific customer?"). Finally, based on a matrix matching item and customer attributes, the algorithm produces a personalized product ranking to answer the original search query ("I predict that Tripod A will be most appealing to this customer, followed by Tripod B, and finally Tripod C"). In what follows are the technical specifications of the algorithmic model in the different stages.

The Informational Retrieval System

First, when processing a user's search query, the *Okapi BM25* algorithm (Robertson & Zaragoza, 2009) is applied to retrieve items whose attributes match the given input string. This step serves to subset a number of potentially relevant items from the overall pool of items available on a website. The algorithm is a bag-of-words information retrieval function that uses an item's attributes to compute its relevancy score for an input string based on the combination of different frequency-based measures. Subsequently, *LambdaMART* (Burges, 2010), a supervised machine learning model developed by Microsoft, is applied on the subset obtained in order to further identify and rank items based on their relevancy for the specific search query.

The original algorithm is modified with *Logged Bandit Feedback Guided Learning* proposed by Swaminathan and Joachims (2015), and the approach of Wang et al. (2018) to overcome *position bias*. Position bias is the tendency of customers to click more frequently on items that appear on the top than on the bottom of search outputs, regardless of their relevance (Wang et al., 2018). Position bias therefore represents an inherent challenge for all click data-based item rankings and needs to be addressed with the aforementioned adjustments.

Stage 1, the information retrieval system, aims to identify and rank each product based on its degree of relevancy for a customer's search input. This stage does not yet create the final product ranking based on customer preferences. Only in the following stage, the recommender system, are customer attributes used to create personalized recommendations from the retrieved products.

The Recommender System

The core method used for building recommendations in the particular algorithmic model used in this study is called *matrix factorization*. Matrix factorization falls into the category of *collaborative filtering* algorithms, which leverage information about the preferences of an entire customer base to produce product recommendations (Ricci et al., 2010). The algorithm relies on input information about customers which can be composed of explicit feedback, such as product ratings, or

(*Appendices continue*)

implicit feedback, such as demographic data or browsing and purchase patterns.

The algorithm takes this customer input information and matches it with product input information to form a user-item interaction matrix, and learns its recommendations via the *element-wise Alternating Least Squares (eALS)* technique (He et al., 2016; Koren et al., 2009). The output of this model (i.e., user–product associations) is then used to personalize the product ranking of the customer's choice set output.

Appendix B

Equivalence Tests

Equivalence tests check whether an estimate and its uncertainty (i.e., confidence interval) falls within a region of "practical equivalence" (*ROPE*). If the estimate and its confidence intervals falls within this region, this is considered to be equivalent to the null. I conducted the analysis following the "TOST rule," that is, a Two One-Sided Test procedure (Lakens, 2017). According to this rule, practical equivalence is rejected when the narrow confidence intervals include or exceed ROPE boundaries and accepted when they are completely inside the ROPE. Otherwise the decision whether to accept or reject practical equivalence is undecided. The analysis was conducted using the bayestestR package in R (Makowski et al., 2019). See summary of the results in Table B1.

Table B1
Equivalence Tests for Models 1–4

Effect	95% CI *LL*	95% CI *UL*	SGPV	Equivalence	*p*
		Model 1: Purchase			
Retailer A					
Intercept	−11.26	−11.25	<.001	Rejected	>.999
Segment (LA)	−0.39	0.68	.339	Undecided	.612
Retailer B					
Intercept	−13.01	−13.01	<.001	Rejected	>.999
Segment (LA)	−0.42	0.89	.277	Undecided	.7
		Model 2: Add-to-cart			
Retailer A					
Intercept	−10.08	−9.18	<.001	Rejected	>.999
Segment (LA)	−0.46	0.47	.389	Undecided	.522
Retailer B					
Intercept	NA	NA	NA	NA	NA
Segment (LA)	NA	NA	NA	NA	NA
		Model 3: Products viewed			
Retailer A					
Intercept	−0.39	−0.32	<.001	Rejected	>.999
Segment (LA)	−0.04	0.06	>.999	Accepted	.002
Retailer B					
Intercept	−0.75	−0.71	<.001	Rejected	>.999
Segment (LA)	−0.02	0.04	>.999	Accepted	<.001
		Model 4: Session length			
Retailer A					
Intercept	96.57	106.72	<.001	Rejected	>.999
Segment (LA)	−5.09	9.21	>.999	Accepted	<.001
Retailer B					
Intercept	21.75	23.2	<.001	Rejected	>.999
Segment (LA)	−1.49	0.57	>.999	Accepted	<.001

Note. Equivalence test with 95% CI under the TOST rule for Models 1–4. SGPV is second generation *p* value and represents the proportion of the confidence interval range that is inside the ROPE. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; LA = low attractiveness; NA = not applicable; ROPE = region of practical significance; TOST = Two One-Sided Test.

(*Appendices continue*)

Appendix C

Analysis of Click and Purchase Distribution by Condition

To understand if there was a difference between conditions in the distribution of product clicks or purchases across the item list, the following analysis was conducted.[C1] First, new variables were coded to capture, for each observation, (1) the number of product clicks that came from the top 1–7 items (henceforth "top click"), (2) the number of product clicks that came from items 8 and onwards (henceforth "bottom click"), (3) the number of product purchases that came from the top 1–7 items (henceforth "top purchase"), (4) the number of product purchases that came from Items 8 and onwards (henceforth "bottom purchase"). This analysis was conducted in Retailer A only, since Retailer B did not provide click and purchase position data. See Table C1 for descriptive statistics.

Similarly to the main analysis reported in the article, multilevel models were applied with random effects accounting for between participants variation. All models were multilevel negative binomial regression estimated using maximum likelihood (ML) estimation approach. No model uncovered statistically significant differences between the HA and LA conditions, meaning there were no differences between these conditions in the click or purchase distribution. The results are reported in Table C2.

[C1] I thank the reviewer who suggested to include this analysis in the article.

Table C1
Descriptive Statistics for Click and Purchase Distribution by Condition

Metric	HA	%	LA	%	Total	%
Top click	3,161	43.9	3,245	44.2	6,406	44.1
Bottom click	4,039	56.1	4,093	55.8	8,132	55.9
Total click	7,200	100.0	7,338	100.0	14,538	100.0
Top purchase	102	75.6	102	73.9	204	74.7
Bottom purchase	33	24.4	36	26.1	69	25.3
Total purchase	135	100.0	138	100.0	273	100.0

Note. Retailer B did not provide click and purchase distribution data. Descriptive data given on a search session level and does not account for dependencies in the data. HA = high attractiveness; LA = low attractiveness.

(*Appendices continue*)

Table C2
Multilevel Models for Click and Purchase Distribution

Effect	*B*	Statistic	*p*	*OR*	95% CI *LL*	95% CI *UL*
			Model 1: Top click			
Condition (LA)	0.04	0.76	.449	1.04	0.95	1.14
Intercept	−1.00	0.37	<.001	0.37	0.34	0.37
			Model 2: Bottom click			
Condition (LA)	−0.05	−0.99	.324	0.95	0.87	1.05
Intercept	−1.24	−32.81	<.001	0.29	0.27	0.31
			Model 3: Top purchase			
Condition (LA)	−0.39	−0.82	.411	0.68	0.27	1.71
Intercept	−11.85	−4702.71	<.001	0.00	0.00	0.00
			Model 4: Bottom purchase			
Condition (LA)	0.04	0.04	.966	1.04	0.17	6.56
Intercept	−12.87	−15.50	<.001	0.00	0.00	0.00

Note. All data are from Retailer A. Retailer B did not provide click and purchase distribution data. *OR* = odds ratio; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; LA = low attractiveness.

Received January 16, 2023
Revision received November 21, 2023
Accepted February 16, 2024 ■