The Effectiveness of Nudging in Commercial Settings and Impact on User Trust

Katarzyna Katner
City, University of London
London, UK
katarzyna.katner@city.ac.uk

Radu Jianu
City, University of London
London, UK
radu.jianu@city.ac.uk

ABSTRACT

Persuasive technologies and nudging are increasingly used to shape user behaviors in applications ranging from health and the environment to business. A thorough understanding of the effectiveness of nudges across different contexts and whether they affect user perception of a system is still lacking. We report the results of a controlled, quantitative study with 20 participants which focused on testing the effectiveness of three different nudges in an e-commerce environment and whether their use has an impact on participants' trust. We found that products nudged via an anchoring effect were more frequently "bought" by participants, and that while participants deemed a store version implementing nudges and one which did not to be equally trustworthy, they perceived the former as technically inferior. Overall we found the effects of nudging to be less dominant than reported in previous studies.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Laboratory experiments*; HCI theory, concepts and models.

KEYWORDS

Persuasive technologies; Nudging; Quantitative evaluation; User study.

SIGCHI, May 2019 , Glasgow, UK
2010. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

ACM Reference Format:
Katarzyna Katner and Radu Jianu. 2019. The Effectiveness of Nudging in Commercial Settings and Impact on User Trust. In *Proceedings of ACM Woodstock conference (SIGCHI).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

Thaler and Sunstein popularized the terms *choice architecture* - how choices are presented to consumers - and *libertarian-paternalism* - designing choice architectures that 'nudge' consumers towards beneficial decisions - in the behavioral economics arena [9]. In human computer interaction (HCI), Fogg defined *persuasive technology* as "interactive information technology designed for changing users' attitudes or behavior" and captured its behavioral underpinnings with *Fogg's Behavior Model* (FBM) [2, 3].

Knowledge gaps related to digital nudging:

- Research on persuasive technology in commerce is limited.
- There is a lack of research evaluating the trustworthiness of systems implementing nudges

We contribute:

- A controlled quantitative evaluation of the effectiveness of three nudges in the context of online commerce
- The first controlled study that measures the effectiveness of nudges in conjunction with user trust

INTRODUCTION

As persuasive technologies and nudging are increasingly used to shape user behaviors in health, environmental protection, education, and commerce [4], a robust evaluation and understanding of nudging is important [1]. The study presented here contributes towards this goal in two ways.

First, we add to a relatively limited number of studies investigating digital nudging in commerce. While persuasive technologies and nudging were studied in the context of many application areas, a review by Hamari et al. shows that empirical studies on persuasive technologies focus predominantly on health and well being (48%), and the environment (21%). Conversely, only 6% of studies targeted commercial applications [4]. This finding is echoed in a more recent review by Mirsch et al. who examined 65 published studies related to nudging, libertarian paternalism and behavioral economics [6]. As contextual factors significantly influence user behavior [1], it is important that nudges are studied in different domains and usage contexts.

Second, we contribute one of few evaluations of the trustworthiness of systems implementing nudges in conjunction with measuring the effectiveness of these nudges. Matthews et al.'s systematic review of digital persuasion that promotes physical activity [5] reflects that there is a lack of work measuring the credibility of systems that employ digital nudging. Int the e-commerce domain Djurica and Figl measure customers' attitude towards sites which implement digital nudging and hypothesize that products incorporating time-pressure nudges will more likely be chosen than products that do not have such cues, but that e-commerce sites using nudges to put pressure on customers may be evaluated less favorably than sites that do not.

METHODS

We used within-participant A/B testing of shopping behavior in two mock online grocery stores, one implementing nudges (v1) and one not (v2), to measure the effectiveness of three specific nudges. 20 participants took part in the study conducted over three weeks.

Evaluated context and nudges: We aimed to evaluate nudging in an online commercial setting and chose online grocery shopping as we believed it to be a scenario that many participants could relate to. To reduce the complexity of the study we opted for a specific scenario: shopping for a weekly supply of breakfast foods. Mintel Group Ltd (2018) reports that people's breakfast choices fall broadly

Nudge A: Displaying item popularity has been selected to cater to people's desire to fit within social norms and build on information of others (i.e., "if it's popular it must be good")

Nudge B: Price offers with limited time duration was selected to cater to people's desire to save money and play to the scarcity effect

Nudge C: Price offers with a set maximum quantity per customer acted as an anchor with hopes that people might purchase higher quantities of items

under 11 categories (e.g., cereal, fruit, pastries/baked goods), making it feasible to create test online stores geared towards breakfast essentials that are both realistic and controlled.

Selecting appropriate nudges that suit the environment was crucial for an effective study design. Dolan et al. report effects that are known to be most influential in changing behavior [1]. We used these to inspire three nudges to evaluate in our online grocery shopping context (sidebar left). We targeted nudges that operate at people's automatic level and covered multiple cognitive effects (e.g., scarcity, social norms, anchoring).

Materials: We designed two versions of online breakfast grocery shops, one incorporating all three nudges (v1) and one without nudges (v2). We also considered the option of designing four different versions, one without nudges, and three separate ones for each individual nudge but decided against it so as to reduce the complexity and resource requirements of our experiment. While the effectiveness of nudges could be tested in a single test-store incorporating all nudges, a version without nudges was needed to explore whether participants perceived it as more trustworthy.

To create functional test-stores with a realistic feel we decided to use an existing e-commerce platform (Shopify). Ultimately, our websites consisted of key pages necessary for completing the task: homepage, category pages, product pages and shopping cart page.

We took several measures to simulate a shopping experience that was realistic but controlled enough to isolate nudging effects and reduce confounding factors. We used Mintel's 2018 report on the most popular breakfast products to select 33 products (e.g. bagels, muesli) and offered each of these products at three different price points (Fig. 1). Furthermore, we gathered and averaged real item pricing from popular UK grocery stores (i.e., Sainsbury and Tesco). One of the main things that people look at when making decisions in an e-commerce environment are product pictures (Mintel, 2018). To isolate nudging effects it was important to select product imagery that would not influence participants' decisions . We used Coyne's photography guide to select product pictures that were high quality, had minimal detail, were consistent across our mock inventory, had the same color background, used the same photo style, and had no visible branding on products. This meant that photos of different product price points of the same subcategory were similar enough to avoid bias but distinctive enough so users know they're looking at a different product.

Once the visual designs for nudges were completed, nudges were allocated to products. In order to give the nudged and not nudged products a fair setting for comparison, it was decided that in the version of the store implementing nudging 50% of products would have nudges and 50% would not. Which products would be nudged was decided randomly via a script. The pool of products selected for nudging was then allocated one of the three nudges also at random whilst ensuring that each nudge was represented equally.

Figure 1: For each category of product we offered three specific products with slightly different prices.

Participants: 20 participants took part in our study. The sample size was established based on similar studies in the literature and with consideration for time and financial limitations of the project. Sauro's guide to finding the right sample size was consulted [7].

As we felt it was important that our participants were (or could be) users of an online grocery store we designed a screening questionnaire and used it to select our 20 participants from a pool of 44 candidates. Specifically, we filtered out candidates which were not open to shopping online, who were under 18 or not UK residents (for legal reasons), and who had UX or marketing expertise and could have been familiar with nudging designs.

Procedure: We opted for a within-subjects design (i.e., each participant used both versions of our grocery stores) as we had limited access to participants and wished to capture changes in the behavior of individual participants between the two versions. The order in which the two systems were used was alternated between two halves of our participant pool: a first half was shown the no nudge version in the first round of the study and the second group the nudged version. Then in the second round, the versions were swapped. Participants used the two versions approximately one week apart to reduce learning effects and create a realistic weekly grocery shop. Participants were incentivised by being entered into a draw for a £50 Amazon coupon.

The study was delivered to participants via Loop11, a remote user testing platform. Loop11 enables the design of studies, including tasks and questionnaires, and can collect video and audio data. For the purpose of the current study only the participant's screen was recorded during the sessions. We opted for remote testing so that participants could complete the task in a setting of their choice at a time convenient to them whilst the recruitment was not restricted to a particular geographic location [8].

Before the actual study commenced a pilot study helped to debug the test environment and gather some qualitative information about the interface. The pilot study was first performed by the researcher and then by two participants. Task instructions were found to be clear and the process of completing the exercise was straightforward. Only minor adjustments were needed to be made, such as for example to the phrasing of the post-task questionnaire, and Loop11 account settings.

Data collected: We collected screen-recordings of the participants' activity as captured by Loop11. We later parsed these videos to extract total cart value, number of items purchased (all), number of un-nudged items purchased, number of nudged items purchased, number of items nudges with nudges A, B, or C, and time on task. Additionally, a post-task questionnaire was used to collect participants' self-reported perception of our two stores' technical performance and trustworthiness.

REFERENCES

[1] Paul Dolan, Michael Hallsworth, David Halpern, Dominic King, Robert Metcalfe, and Ivo Vlaev. 2012. Influencing behaviour: The mindspace way. *Journal of Economic Psychology* 33, 1 (2012), 264–277.

[2] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do.

Although 20 participants took part in the study, data from three participants had to be excluded because the participants in question used the same device to complete the study as some of the other participants. This meant that when they were redirected to the study websites, the items from the previous session remained in the cart. We were unable to determine if this influenced their behavior.

RESULTS

We found no statistical difference between participants' preference for nudged vs. un-nudged items, even though overall participants added approximately 17% more nudged items to their shopping carts. However, we found that participants preferred items nudged by nudge A over those nudged by nudge C. Finally, while participants ranked both system versions as equally trustworthy, they ranked the one using nudges as technically inferior.

Overall nudge effectiveness: We used a paired t-test to check whether there was a statistically significant difference between the collective number of nudged items compared to the number of un-nudged items selected by participants. We used only items selected by participants in the session which employed nudging, i.e., in the session in which 50% of items were nudged and 50% were not. We found that even though participants added about 17% more nudged products into their shopping cart, this difference was not statistically significant ($p = 0.23$).

We performed a similar comparison between items purchased in one system (v1 - with nudging) versus the other (v2 - without). We found that the number of items added to the shopping cart was 11% higher in v1 than in v2 but that this difference was also not significant, as revealed by a t-test.

Comparative nudge effectiveness: To determine if some nudges perform better than others, we compared the counts of items that participants added to their carts corresponding to each of the three types of nudges. We found nudge C to be most popular (38 selections), followed by nudge B (29 selections), and nudge A (17 selections). A single factor ANOVA test over the counts of the three nudge groups revealed the differences between groups to be statistically significant ($p = 0.02$).

To verify differences in the mean values of counts between all possible nudge pairs, we conducted three pairwise t-tests and interpreted the results using a Bonferroni correction. We found no statistically significant difference between items counts for nudges A and B, and nudges B and C, but we did find a statistically significant difference between the counts of Nudge A and Nudge C ($p = 0.011$).

Impact on trust: In the post-study survey participants were asked four questions related to trust (e.g., " *Shop name* is a shop I could trust", "I felt that *Shop name* had my best interest in mind").

We asked our participants to assess the technical level our websites along five dimensions (below). This lead to an interesting results (right).

Thinking about the site you just visited, which of the following statements apply to your experience?

- ❑ Overall, this site worked very well technically.
- ❑ Visually, this site resembled other sites I think highly of.
- ❑ This site was simple to navigate.
- ❑ On this site, it was easy to find the information I wanted.
- ❑ This site clearly showed how I can contact or communicate with the vendor
- ❑ None of these apply

Participants answered these using a 5-point Liker scale ranging from strongly agreeing to strongly disagreeing. When aggregating and quantifying the data we found no statistically significant differences between version 1 (nudged) and version 2 (un-nudged).

A surprising result came from asking participants to rank the technical performance of the systems (left). Even though the two test-stores were essentially the same, a paired t-test revealed that participants perceived version 2 (un-nudged) to be technically superior to version 1 (nudged) ($p = 0.04$).

DISCUSSION

Given results reported in previous studies with similar goals and methodologies (e.g., Schneider et al. [10]) we were surprised to find no statistical difference between the number of nudged and number of un-nudged items that participants added to their shopping carts. This may indicate that users are, or are becoming, more immune to nudging than we expect, at least in a commercial setting. However, the result may also be a consequence of limitations in our study's designs, such as for example a low number of participants. Our study did find that customers added to their shopping cart approximately 17% more nudged items than un-nudged items, and although the paired t-test results determined that the difference was not significant, the findings show promise for further research.

An interesting result was that participants didn't find the nudged version of the test-store to be less trustworthy than the un-nudged version, but they perceived it as technically inferior. This was unexpected since the sites' navigation and overall features were identical with the only difference being the presence of nudges. The idea that the use of nudges could negatively impact the perception of a site's technical performance is surprising and worth investigating further.

Our methodology accounted for confounding factors, allowed for multiple nudges to be compared with minimal use of the participant's time, and allowed for comparisons between nudged and un-nudged systems as a whole. Our study could be extended to test additional nudges and future work could include tests with more participants, and more diverse nudges. Different execution of nudges could give rise to different results. As demonstrated in the methods section, there are numerous ways in which a nudge could be executed. Color, size and the content could all have an impact on a nudge's prominence and therefore its effect. In addition to testing a range of nudges, the way of executing each nudge could also be tested.

CONCLUSION

We evaluated three nudges quantitatively in terms of their ability to shape user buying behavior in a mock online store. Unlike previous results, we found nudges to be relatively ineffective in influencing participants' buying patterns, except for a few small effects. We also evaluated users' perception of the trustworthiness of stores employing nudging. Participants rated the store with nudges as equally trustworthy than the one without, but perceived the former to be technically inferior.