Idea evaluation by citizens: The Hamburg Fab City Maker Challenge case

David Pacuku, Johanna Schnier, Tobias Redlich, Manuel Moritz, Marvin Klein, Christina Raasch, Christian Lüthje
Kühne Logistigs University, Technische Universität Hamburg, Helmut-Schmidt-Universität
Hamburg, Germany
david.pacuku@the-klu.org, johanna.schnier@the-klu.org, tobias.redlich@hsu-hh.de, manuel.moritz@hsu-hh.de, marvin.klein@tuhh.de, christina.raasch@the-klu.org, c-luethje@tuhh.de

Abstract - **Experts play an important role in the evaluation of ideas. Owing to their experience and knowledge, they are believed to be best suited to objectively evaluate ideas. At the same time, expert evaluations sometimes fail to accurately reflect stakeholder preferences. For this reason, more and more organizations have begun to solicit 'non-expert' evaluations from the 'crowd' (e.g., via community voting). We illustrate how organizations can involve stakeholders in idea evaluation in the context of the Fab City project. Integrating citizens into the innovation process is particularly interesting in a Fab City–a city that produces many of the products it needs locally–since it enables one to develop regional preferences into tailored solutions. Here, citizens evaluate a solution, implement it as a local product, and ultimately share the design (mostly open-source) within the community. In the course of a Fab City initiative in Hamburg–the Maker Challenge–citizens were invited to evaluate almost 100 innovative ideas in pairwise comparisons. We draw on over 20,000 votes from around 400 citizen judges to assess the crowd opinion formation process. Unlike in other studies, our data allow us to measure the time every citizen judge needed to make their decision. Thus, we can conclude how diligent citizen judges are and, conversely, how much attention an idea can attract. Looking at the citizen judges' evaluation times, it appears that, in most cases, the opinion-formation process is spontaneous. In a few cases, however, a substantial amount of time is spent. All the evaluation times followed a Poisson distribution. Further, we compare expert evaluations with crowd evaluations and relate the results to the literature. We conclude that the crowd evaluation process resembles the expert evaluation process. Also, we show that citizens optimize their cognitive effort over time. For real-world cases, such as the Fab City project, this implies that not only the ideation process and the implementation process can be carried out by local citizens, but also that the evaluation can be done within a community without significant loss of quality, but with much lower effort on the part of organizers.**

Keyword - Fab City–Maker Challenge, Citizen Innovation, Community Voting, Evaluation Process

I. Introduction Section

The Fab City global initiative was established in the 2010s with the aim to challenge cities whose current consumption and resource use are unsustainable. Currently, around 40 cooperating cities and regions have committed to become more self-sufficient and, therefore, more sustainable. There are different focus areas, such as food, energy, materials, waste, water, and information systems. We focus on the Maker Challenge, a local competition that invited citizens to share innovative ideas that could possibly be produced locally in the future. These ideas were then evaluated by the local population. Specifically, in Hamburg in June 2022, almost 100 innovative ideas were collected, of which the most popular 20 were implemented as prototypes with the help of experts. During this event, we collected more than 20,000 data points. The amount of data enabled us to describe the ways crowd members and experts vote on both an individual and an aggregated level.

By comparing citizens to experts, we conclude that the ways both groups vote are very similar. Coupled with the literature, we conclude that evaluation by the crowd is an appropriate substitute for evaluation by experts. Further, the Maker Challenge revealed the optimization of cognitive effort by citizens during the evaluation process.

The following chapter sheds light on the current literature and research frontiers, and forms the basis for our hypotheses, which we elaborate in chapter 3. In chapter 4, we analyze our findings, discussing them in chapter 5.

II. Literature Review

The option to outsource the innovation process to others goes hand in hand with a great many people being connected to the global network that is the Internet [1], [2]. Under the neologism *crowdsourcing*, Howe (2006) [3] describes how companies have begun to collect innovative ideas in web-based ways through consumers. These ideas are then evaluated by the crowd [4]. The practical idea behind this lies in the fact that the sheer volume of ideas collected simply cannot be handled internally without great effort. Trials to evaluate these ideas, using the resources of a firm, usually ended up being both very inefficient and time-intensive. For instance, it took Google no less than 3,000 employees and three years to evaluate 150,000 ideas submitted to its Project 10 to the 100 until the first proposals could be implemented [5]. IBM once even employed 50 senior executives for several weeks to evaluate 50,000 ideas [6].

Internal idea evaluation has not only proved inefficient, but also bears the risk of being biased. For instance, Asplund et al. (2021) [7] found that internally employed experts are biased toward exploitable ideas, owing to risk-aversion, which comes with lower uncertainty about new, innovative ideas and a preference for returns that are closer (rather than further away) in time.

For a city that has self-sustainability and resource efficiency as its goals, the possibility of outsourcing the evaluation process in the spirit of crowdsourcing seems comparable to having this evaluation done by experts [8]. The term *wisdom of the crowd* is often encountered here, defined as an observed statistical phenomenon characterized by the aggregated opinion of a population being closer to a true value than most individual evaluations, estimating for instance heights or weights [9] even though few individual votes diverge very far from the truth. Studies show a significant agreement between the crowd's opinions and experts' ratings [10] as well as overall concurrence in the relative rankings of ideas between the crowd and experts [11]. Further, Kornish

and Ulrich (2014) [12] concluded that online consumer panels are even a better way to determine what is a good innovative idea than expert ratings.

Kahneman and Tversky (1981) [13] elaborated on individual errors and biases in judgment under uncertainty and define *intuition* by using it in three ways. One, an analysis can be called intuitive if it does not rely on an analytical method or a deliberate calculation. Two, a rule or a fact can be considered to be intuitive "if it is compatible with our lay model of the world." Three, a procedure is intuitive if it is integrated into our daily living.

It is crucial to define intuition when describing the ways in which both citizens and experts evaluate innovative ideas. At some point, however, the demand for high concentration lead to effects such as cognitive overload [14]. It is also possible that participants reach their computational limits [15] and, owing to a lack of concentration, make performance errors, which limit individuals' assessments of ideas. As a result, persons make mental abbreviations, which leads to individual biases when assessing innovative ideas.

Kahneman's *Thinking, Fast and Slow* [16] adapts the idea of differentiated ways of thinking in a similar way. He uses the metaphor of two systems of thinking that co-exist in persons' minds. The biggest difference between these two systems is the ways in which they are used to solve tasks or take decisions. System 1, which is the one we use most often, follows heuristic rules, and is biased, since it is characterized by an intuitive, automatic, unconscious, and effortless way of deciding. System 2, which we seldom use, is a more 'rational' way of thinking. It closely assesses the context and is slow, controlled, effortful, and statistical. While system 1 is mostly used to solve easy tasks in daily life, system 2 is used to solve problems in an enduring way.

Using this as a theoretical foundation, and considering the evaluation times, we hypothesize that when both experts and citizens use two different systems of thinking, this would be recognizable as a pattern during evaluation processes. Further, if this pattern is indistinguishable between citizens and experts, one can conclude that their evaluation processes are similar. This supposition is further underlined by considering the aggregated duration, where we expect to obtain comparable Poisson distributions, plotting the evaluation times against their relative frequencies (i.e., the violin density).

Further, we consider Verplanken, Aarts, and Knippenberg (1997) [17], who researched habit-based decision-making. Previous decisions that led to reasonable or positive outcomes can lead to similar behaviors, especially concerning future decisions. Further, Hoeffler and Ariely (1999) [18] concluded that this occurred most rapidly when participants had to decide between two options. This behavior can be explained by the participants' urge to maximize the efficiency of their cognitive effort. Ariely and Zakay (2001) [19] summarized that this *preference consolidation* occurs as participants are forced to consolidate how they feel about characteristics of these ideas and forces them to stabilize these thoughts as they face tradeoffs between different ideas. Ariely, Loewenstein, and Prelec (2000) [20] drew parallels to the herding effect and call this effect–when participants internally simplify the evaluation process by using previous decisions' characteristics–*the self-herding effect*. Taking mental abbreviations to minimize cognitive effort leads to swifter decisions if more evaluations have been carried out prior to the current one. Thus, we hypothesize that, on average, individual voters increase the evaluation process speed with the number of votes made prior.

Keeping in mind the potential susceptibilities to distortions of the wisdom of the crowd, we argue that our experimental setup minimizes potential intrusions that could lead to strongly biased voting behavior. Arguing for an unbiased dataset, we focus on the following gap in the research, asking whether, both at the individual and the aggregated levels, the ways in which citizens and experts evaluate innovative ideas are similar.

III. Methodology

Using quantitative analysis techniques, we focus on how citizens evaluate innovative ideas, so as to draw comparisons to the way experts evaluate them. We use descriptive statistics and, by depicting the dataset, we can describe the way the crowd evaluated the ideas; second, we can compare the two groups.

The data were collected in June 2022 through a unique innovation competition in Hamburg–the Fab City Maker Challenge. The challenge had three decisive steps.

First, in the spirit of a social, sustainable, innovative, and self-sufficient community, citizens were asked to share their innovative ideas by uploading them to the Challenge's online platform. Incentives were created for the best ideas. The top 20 ideas would participate in a Prototyping Workshop, and the best idea would win a 3-D printer valued at 400 EUR. Further, more intrinsic reasons for citizens to participate were to gain publicity for their ideas through the Fab City Hamburg Maker Expo. Finally, participants could also take advantage of the community feedback in the form of comments and could improve their ideas accordingly. Overall, almost 100 ideas were submitted by a diverse group of people living in Hamburg.

To evaluate all the ideas in our campaign, we asked citizens from the broader Hamburg region to evaluate them online. The evaluation process followed the Swiss System tournament. In round 1, two innovative ideas were randomly assigned to compete against each other. The pairwise comparisons were randomly assigned to the citizen judges, who repeatedly indicated which idea they preferred or whether they considered both to be equally good. In every comparison, each idea received three points if it won, two if there was a tie, and zero if it lost. The voting system was open for about a week. Almost 20,000 evaluations were done by 400 citizens. The dataset was unique owing to the amount of information we could collect from each evaluation. By considering the anonymized data, we could retrace the duration between two evaluations, which enabled us to further characterize how the evaluation was done.

The top five ideas from the citizen evaluation process were immediately implemented in the Prototyping Workshop, while the remaining 15 (in the top 20) were distributed to experts for evaluation. The panel of experts comprised seven experts, who evaluated the remaining ideas in the top 50. In this process, 350 expert evaluations were generated.

As the citizens and the experts did not evaluate exactly the same group of ideas, we were unable to compare the respective outcomes of their evaluations of the two ideas. As per the

literature, we do not expect evaluations by citizens and experts to differ significantly. Further, this set-up excluded biasing effects, such as the herding effect [21]. We focus on the ways these two groups performed in the respective tasks to evaluate these ideas.

We considered and compared, at an individual and an aggregated level, with descriptive statistics, the data we collected for both groups. Specifically, we drew the observable densities for the evaluation time for each individual and the distribution on an aggregated level. By comparing these graphics, we could then draw a conclusion about the comparability of the two groups, which differ significantly in their knowledge, experience, and expertise.

By considering the average evaluation time, conditioned on its absolute position in the rank of the consecutive evaluations, we could check whether the evaluation time decreased by rank.

IV. Results

To analyze the judges' diligence in evaluating the ideas, we compared the evaluation times across the citizens. By taking the difference in time when citizen judges were exposed to a new idea and when they judged them, we could measure how much time a judge needed to evaluate it. An evaluation time was attributed to each judge and, since each judge voted multiple times, we could derive a distribution of duration for each judge. We ordered the citizen judges according to their average evaluation time, and then put them in 20 groups with the same number of judges. Next, we plotted the duration by means of violin densities to graphically illustrate the relative evaluation duration frequency. That is, the more surface a density took within a given time, the more frequently this observation has been made, see Figure 1 and Figure 2. We ran the same analysis for the expert judges Figure 3.

By comparing the outcomes, we drew the following conclusions: A clear pattern can be observed independently of the average evaluation time of the citizen judges; it is even observed for the expert judges: Usually, judges attempt to evaluate ideas spontaneously. Comparable to Kahneman's two systems of thinking [16], there is a pattern: the judges used their intuition to evaluate most of the ideas. This is surprising, since the two ideas given to the judges were evaluated relatively similar. In other words, even though we considered the two ideas to be very similar in quality, the judges were able to swiftly decide which one they preferred. This happened most of the time, as can be seen by the fat tails of the violin distribution at the lower end.

Further, some evaluations were not made spontaneously, as judges invested more time in evaluating some ideas. For the cases that judges did not prefer one idea over another, we provided the tie option. Thus, we could exclude that judges took longer to evaluate because they were indecisive on how to evaluate an idea. The evaluation time can therefore be considered as a measurement of the judges' diligence.

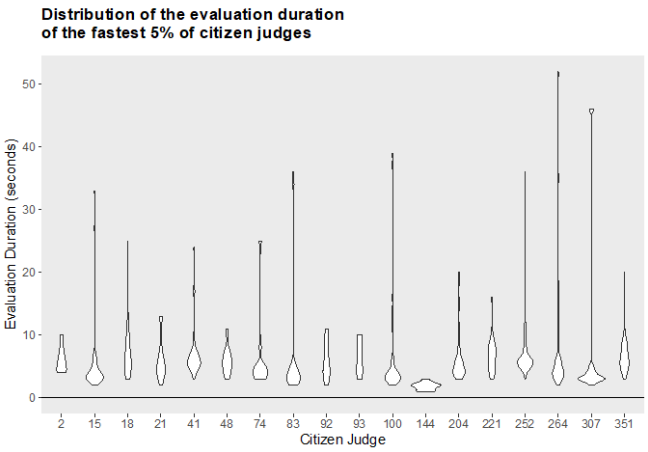

Figure 1: Distribution of the evaluation times of the fastest 5 % of the citizen judges.

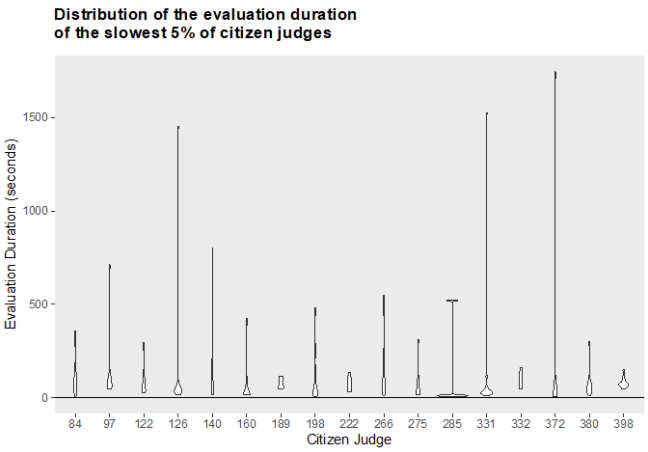

Figure 2: Distribution of the evaluation times of the slowest 5 % of the citizen judges.

When comparing citizen judges to the experts, there was no difference in the patterns, in the sense that judges invested more time to evaluate some idea pairs. For most ideas, however, the evaluation time was short. Thus, the same pattern can be observed for the citizens and the experts.

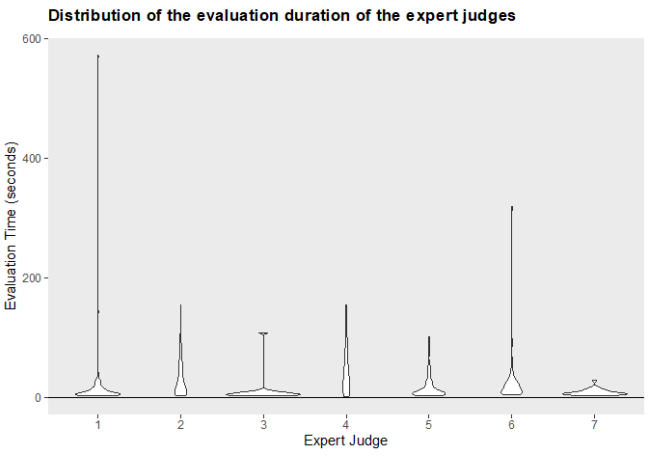

Figure 3: Distribution of the expert judges' evaluations

We can observe comparable results when considering the aggregated evaluation times of the citizens, see FIGURE 4 and the experts FIGURE 5. Specifically, we obtained two comparable Poisson distributions. Both groups had a comparable evaluation time, with a mean of around 21 seconds for the citizens and around 22 seconds for the experts, which are depicted by the blue vertical dash-lines in FIGURE 4 and FIGURE 5.

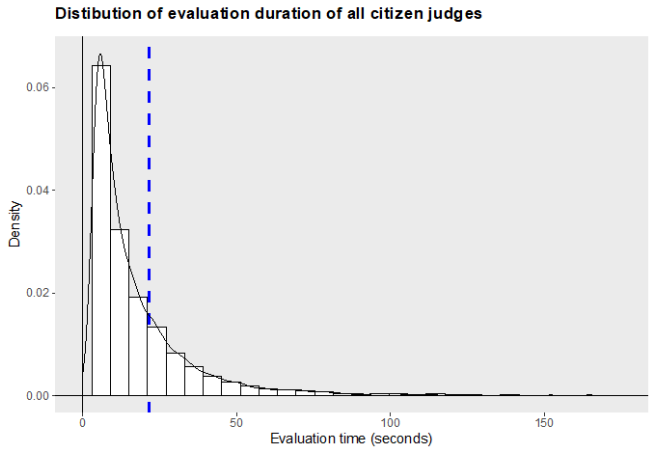

FIGURE 4: DISTRIBUTION OF THE EVALUATION TIMES OF ALL THE CITIZEN JUDGES.

As we consider the average evaluation time, given the respective rank of the evaluation, we concluded that a rank higher than 50 would lead to biased results owing to fact that we had fewer observations per rank. Thus, we excluded the evaluations ranked higher than 50 from this analysis. In FIGURE 6, one can immediately see the relationship between the ranking and the average evaluation time. The red dash-line depicts the resulting line in an OLS estimation, and the blue shadow the respective confidence interval. Driven by habit-based decisions and the intrinsic unintentional urge to maximize the efficiency of cognitive effort, one can clearly conclude that the habitus (socially ingrained habits, skills, and dispositions) reflected by the evaluation ranking plays a direct role in the evaluation time.

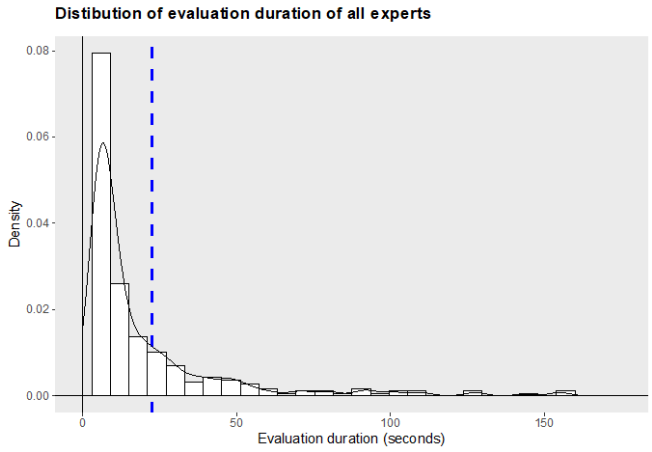

FIGURE 5: DISTRIBUTIONS OF THE EVALUATION TIMES OF ALL THE EXPERTS.

Further, notably, individual traits of both the innovative ideas and the judges played a subordinate role to no role in this analysis. Specifically, we cancelled out individual effects by

taking the average of these observations, making them comparable. This allowed us to run an easy regression, with the ranking in the evaluation frequency as the independent variable and the average evaluation time as the independent variable.

It turned out that, on average, one more ranking led to a reduced evaluation time of 0.369 seconds, which is statistically significant on all conventional significance levels. This means that, for our observation of 50 ranks, the average evaluation time decreased to around 18.5 seconds.

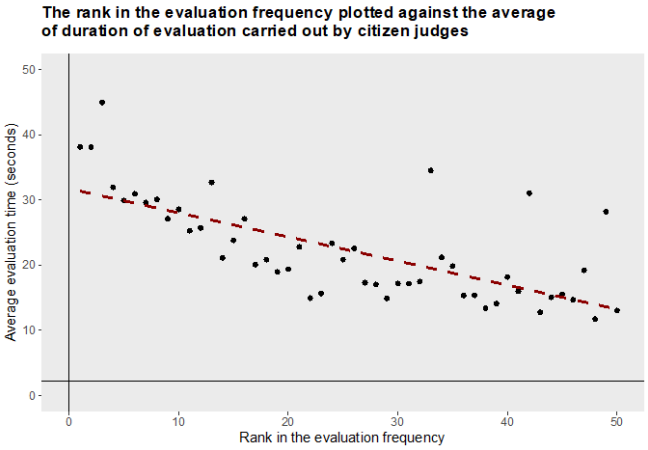

FIGURE 6: THE RANK IN THE EVALUATION FREQUENCY PLOTTED AGAINST THE CITIZEN JUDGES' AVERAGE EVALUATION TIME.

V. CONCLUSION

The idea evaluations of the Maker Challenge ideation contest provide interesting insights into the ways citizens and experts evaluate ideas. Specifically, it seems that both experts and citizens think in two different systems: spontaneous thinking and reflective thinking. The spontaneous thinking system is predominant since it reduces the cognitive effort of the evaluation process. When using the reflective system, which rarely occurs, judges–whether citizens or experts–take more time to make an evaluation. Further, aggregating the two groups' evaluation times led to almost indistinguishable Poisson distributions–another commonality between the evaluation processes of the citizens and the experts, yet only on an aggregated level. Further, looking at the citizens' average evaluation time, ranking plays an important role. The negative correlation between the rank in the evaluation process and the evaluation time can be explained by the reduced cognitive effort.

Since the comparison between evaluations by experts and citizens has been widely explored in the literature, we examined their evaluation processes. It turns out that there are strong similarities between their evaluations processes. The evaluation efforts between the two groups were not distinguishable. This confirms the feasibility of citizen-driven innovation evaluation, as fostered by Fab Cities.

ACKNOWLEDGEMENT

This paper was funded by dtec.bw (the Digitalization and Technology Research Center of the Bundeswehr), which we gratefully acknowledge [project Fab City].

References

[1] Whelan, Eoin, et al., The role of information systems in enabling open innovation., Bd. 15.11 , Journal of the Association for Information Systems , 2014.

[2] Gregori, Enrico, et al., „Smartphone-based crowdsourcing for estimating the bottleneck capacity in wireless networks," *Journal of Network and Computer Applications,* Bd. 64, pp. 62-75, 2016.

[3] J. Howe, „The rise of crowdsourcing," *Wired magazine ,* pp. 1-4, 2006.

[4] Behrend, Tara S., et al., „The viability of crowdsourcing for survey research," *Behavior Research Methods,* Bd. 43.3 , pp. 800-813, 2011.

[5] Blohm, Ivo, Jan Marco Leimeister, and Helmut Krcmar, „Crowdsourcing: how to benefit from (too) many great ideas," *MIS Quarterly Executive,* Bd. 12.4 , pp. 199-211, 2013.

[6] Bjelland, Osvald M., and Robert Chapman Wood, „An inside view of IBM's' Innovation Jam'.," *MIT Sloan Management Review,* Bd. 50.1 , 2008.

[7] Asplund, Fredrik, Jennie Björk, and Mats Magnusson, „"Knowing too much? On bias due to domain‐specific knowledge in internal crowdsourcing for explorative ideas.", " *R&D Management,* Bd. 52.4, pp. 720-734, 2022.

[8] Borda, Ann, and Jonathan P. Bowen, „Smart cities and digital culture: Models of innovation," *Museums and digital culture,* pp. 523-549, 2019 .

[9] Mavrodiev, Pavlin, Claudio J. Tessone, and Frank Schweitzer, „Effects of social influence on the wisdom of crowds," 2012.

[10] Mollick, Ethan, and Ramana Nanda, „"Wisdom or madness? Comparing crowds with expert evaluation in funding the arts.","*Management science,* Bd. 62.6 , pp. 1533-1553, 2016.

[11] Magnusson, Peter R., Johan Netz, and Erik Wästlund, „"Exploring holistic intuitive idea screening in the light of formal criteria.","*Technovation,* Bde. %1 von %234.5-6 , pp. 315-326, 2014.

[12] Kornish, Laura J., and Karl T. Ulrich, „"The importance of the raw idea in innovation: Testing the sow's ear hypothesis.","*Journal of Marketing Research,* Bd. 51.1 , pp. 14-26, 2014.

[13] Kahneman, Daniel, and Amos Tversky., The simulation heuristic, Stanford Univ CA Dept of Psychology, 1981.

[14] C. a. C. S. Benz, „Nudged to Unload: Applying Choice Architecture to Prevent Cognitive Overload of Participants in Open Idea Evaluation," 2019.

[15] P. A. Klaczynski, „Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding.," Bd. 665, 2014.

[16] D. Kahneman, Thinking, Fast and Slow, New York, 2001.

[17] Verplanken, Bas, Henk Aarts, and Ad Van Knippenberg, „Habit, information acquisition, and the process of making travel mode choices," *European journal of social psychology,* Bd. 27.5 , pp. 539-560, 1997.

[18] S. a. D. A. Hoeffler, „Constructing stable preferences: A look into dimensions of experience and their impact on preference stability," *Journal of consumer psychology ,* Bd. 8.2 , pp. 113-139, 1999.

[19] Ariely, Dan, and Dan Zakay, „A timely account of the role of duration in decision making," *Acta psychologica,* Bd. 108.2 , pp. 187-207, 2001.

[20] Ariely, Dan, George Loewenstein, and Drazen Prelec, „Coherent arbitrariness: Duration-sensitive pricing of hedonic stimuli around an arbitrary anchor," 2000.

[21] Hofstetter, Reto, Suleiman Aryobsei, and Andreas Herrmann, „Should you really produce what consumers like online? Empirical evidence for reciprocal voting in open innovation contests," *Journal of Product Innovation Management,* Nr. 35.2 , pp. 209-229, 2018.