Please quote as: Riedl, C.; Blohm, I.; Leimeister, J. M. & Krcmar, H. (2013): The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities. In: International Journal of Electronic Commerce (IJEC), Ausgabe/Number: 17(3), Erscheinungsjahr/Year: 2013. Seiten/Pages: 7-36.

The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities[1]

Christoph Riedl,[2] Ivo Blohm,[3] Jan Marco Leimeister,[4] Helmut Krcmar[5]

PREPRINT

Forthcoming in *International Journal of Electronic Commerce*

ABSTRACT: Given the rise of Internet, consumers increasingly engage in co-creating products and services. Whereas most co-creation research deals with various aspects of generating user-generated content, this study addresses designing ratings scales for evaluating such content. In detail, we analyze functional and perceptional aspects of two frequently used rating scales in online innovation communities. Using a multi-method approach, our experimental results show that a multi-criteria scale leads to higher decision quality of users than a single-criteria scale, that idea elaboration (i.e., idea length) negatively moderates this effect such that the single-criteria rating scale outperforms the multi-criteria scale for long ideas, and finally that the multi-criteria scale leads to more favorable user attitudes towards the website. Based on our experimental data, we applied a bootstrap-based Monte Carlo simulation for assuring robustness of our results. We find that around 20 user ratings per idea are sufficient for creating stable idea rankings and that a combination of both rating scales leads to a 63% performance improvement over the single-criteria rating scale and 16% over the multi-criteria rating scale. Our work contributes to co-creation research by offering insights as to how the interaction of the technology being used (i.e., rating scale), and attributes of the rating object affects two central outcome measures: the effectiveness of the rating in terms of decision quality of its users and the perception of the scale by is users as a predictor of future use.

KEY WORDS AND PHRASES: Website, online innovation community, rating, scale, decision quality, Monte Carlo, bootstrap, user-generated content, co-creation

---

[1] The authors thank the editor and two anonymous reviewers for their helpful comments. This research received funding through the GENIE project by the German Ministry of Research and Education (BMBF) under contract No. FKZ 01FM07027. CR acknowledges support from the German Research Foundation under grant code RI 2185/1-1. All mistakes remain the authors' own.
[2] Harvard University, Institute for Quantitative Social Science, 1737 Cambridge St., Cambridge, MA, 02138, USA, criedl@iq.harvard.edu
[3] Institute of Information Management (IWI HSG), University of St.Gallen, Müller-Friedberg-Str. 8, 9000 St. Gallen, Switzerland, ivo.blohm@unisg.ch
[4] Universität Kassel, Chair for Information Systems, Nora-Platiel-Straße 4, 34127 Kassel, Germany, leimeister@uni-kassel.de; Institute of Information Management (IWI HSG), University of St.Gallen, Müller-Friedberg-Str. 8, 9000 St. Gallen, Switzerland, janmarco.leimeister@unisg.ch
[5] Technische Universität München, Chair for Information Systems (I17), Boltzmannstr. 3 , 85748 Garching b. München, Germany, krcmar@in.tum.de

1. Introduction

Given today's proliferation of social participation and co-creation in e-commerce and Web 2.0 applications [17, 49, 60, 84], mechanisms for user ratings are now found on almost every website. These mechanisms are used to assess rating objects ranging from books (Amazon.com), appliances and consumer electronics (Bizrate.com), movies (imdb.com), to travel services (tripadvisor.com), and, virtually every imaginable category of user-generated content (e.g., YouTube.com).

While ratings of products, services, and online content serve as recommendations towards other users [14, 17, 37, 49, 60], ratings in the context of online innovation communities reflect a proxy measure of idea quality [62]. Online innovation communities invite external actors, in particular end-users, to freely reveal innovative ideas [78, 83]. Through these websites, community members contribute their ideas to be reviewed, discussed, and rated by the user community [27]. In these communities, rating mechanisms help the host organization to filter out the best ideas in order to incorporate them into new/improved products and services [22, 31]. The goal is to identify the 'best' ideas from the viewpoint of the adopting organization. Despite the widespread use of rating mechanisms on websites, they vary in sophistication and features; some provide only basic rating functionality while others use elaborate multi-criteria rating scales [14, 20, 63]. Additionally, high differences regarding the quality and the properties of different rating objects, i.e., ideas in online innovation communities, can be recognized [6] such that different rating scales might be applicable for different rating objects. However, there are still major concerns regarding the design of these rating scales to facilitate effective and efficient decision support for organizations [7].

We use a web experiment and a simulation-based analysis to address the important questions of how functional and perceptional aspects are affected by the design of rating scales used in online innovation communities. Specifically, we ask:

1. How do traits of the rating object, i.e., the degree of elaboration of the ideas, influence the appropriateness of different rating scales for idea evaluation? There is only a limited body of literature investigating functional aspects of website rating scales and no clear design guidelines exist [62]. In particular, it is not entirely clear how elaborate these rating scales should be in terms of rating criteria (single-criteria vs. multi-criteria scales) and if the choice

of rating scale should be based on attributes of the rating object, i.e., ideas, such as the length of its textual representation.

2. How do the different rating scales affect users' attitudes towards a website? User attitudes have been found to be a central predictor for the continuous usage of website [42, 76]. We know that participation in online communities is fluctuating such that creating positive user attitudes is important for gathering a sufficient number of ratings. Besides, positive user attitudes deriving from creative tasks like idea evaluation may favor goals of online innovation communities beyond effective rating and improving innovative strength: building customer loyalty, improving brand awareness, or recruiting new employees [28, 60]. However, perceptional aspects of rating scale usage have merely been addressed in existing research.
3. How many user ratings per rating object (i.e., idea) are necessary for stable rankings? A consequence of dynamic participation in online communities is that not all rating objects receive the same number of ratings. However, to date there is no work that explores the effect of unevenly distributed ratings and the necessary number of user ratings for stable rankings.

Using a multi-method approach, the research goal of this paper is to develop and test a theoretical model of the effects of two different rating scales on decision quality including the moderating effect of idea elaboration as well as user attitudes towards the website. In this regard, we distinguish between a single-criteria rating scale on which all information has to be integrated into a judgment of a single criterion (e.g., idea quality), and a multi-criteria rating scale which consist of several criteria (e.g., novelty and feasibility) for evaluating the rating object. The multi-criteria scale refers to the measurement of a multi-dimensional construct (e.g., idea quality) through a single-item per dimension. Though both rating scales are frequently used online, website designers and scholars usually imply that all rating scales lead to comparable results.

This work provides a comprehensive study that does not simply replicate earlier findings on the psychometric properties of different rating scales in an online context but assumes a more holistic view by considering idea elaboration (a central attribute of the rating object) as a moderator of the relationship between the rating scale and the decision quality of the rating scale user. The study of idea elaboration is motivated by the observation that user-generated ideas vary significantly regarding their degree of detail [6, 22]. Consequently, certain rating scales might not be appropriate for ideas of a certain degrees of elaboration. However, the moderating effect of idea elaboration has not yet been examined, although previous research implies that different rating scales are apt for the evaluation rating objects of varying characteristics, such as the degree of elaboration.

In online innovation communities, single ideas are rated by individual users with an idiosyncratic attitude towards the rating scale and the website. However, companies judge the quality of ideas not based on individual ratings. They are mostly interested in overall rankings of all rated objects to pick the most promising candidates [31] as these rankings attenuate individual decision, and non-systematic measurement errors [37]. We use this aggregated level of analysis to present the overall results from the experiment as well as sensitivity analyses using a Monte Carlo and bootstrap-based simulation to assess when and how stable, aggregated rankings can be constructed out of a pool of individual ratings. For the analysis of the effects of the rating scales on the attitude of rating scale users, the survey data on the user level is analyzed.

One central design decision of our empirical approach relates to the availability of rating information to rating scale users. Only if the collected ratings are independent of each other, can decision quality of users be measured without introducing confounding effects through biasing cross influences between rating scale users [52]. Consequently, the ratings provided by other users were not displayed on the website in our experiment. While we acknowledge that this decision lead to a slight deviation from real-world websites in which user rating would generally be visible, we found it more important to focus on the main condition of interest – the effects of the design of a particular website feature – without introducing additional confounding effects such as social influence and information cascades. As an anticipated consequence, this decision should improve the rating accuracy as confounding influences are removed that were found to diminish decision quality of individuals [cf. 24, 52].

In summary, our research makes the following contributions:

1. From a theoretical perspective, we develop and test a model to analyze the influence of single- and multi-criteria rating scales on users' decision quality and attitude including the moderating effect of idea elaboration. Our results provide a holistic, empirical analysis taking functional and perceptional aspects into consideration.
2. From a methodological perspective, the research uses multiple data sources (system captured experiment data, perceptually anchored questionnaire data, and independent expert ratings)

analyzed using different analytical, statistical, and simulation-based methods to study and interpret the effectiveness of two different rating scales.

3. From a practical perspective, our research provides actionable design guidelines, which could improve the effectiveness of rating scales in online innovation communities. The experimental nature of our research enables us to directly contribute by providing evidence on the design elements that do or do not hold up in practice.

2. Theory and Hypotheses

There is broad consensus among creativity researchers that experts in a given domain are the most appropriate people to evaluate the quality of creative products [1]. This holds true for innovation management where new ideas are generally evaluated by a small team of interdisciplinary experts [1, 6, 22, 31]. However, in regard to online innovation communities where thousands of ideas get submitted, experts are a scarce resource for idea evaluation. Thus, online innovation communities usually offer rating scales for idea evaluation with which community members can help the host of the community to identify the most attractive ideas. From the hosts' perspective, the aggregated user ratings may lead to a reduction of the amount of ideas that have to be reviewed internally by the experts or even to an entire replacement of the expert evaluation. The applied rating scales highly determine the rating decisions of the rating scale users that can be considered as valuable when they help to identify ideas that are considered most promising by the host of the online innovation community, or more specifically its experts. In this regard, existing research defines decision quality most prevalently as the 'correctness' or 'goodness' of decisions [21]. In accordance with this line of research, we define decision quality of rating scale users as judgmental fit between the decision of rating scale users and the objective standard of the independent expert assessment. In the absence of a gold-standard 'true' quality assessment, such an expert-based measure is commonly accepted and widely used in research [15, 39].

2.1. Functional Aspects of Rating Scales

Traditional psychometric theory suggests that scales with multiple criteria produce more valid and reliable measurements than scales with a single criterion. Multiple response criteria lead to more consistent results as they are less susceptible to response biases of respondents [56]. Moreover, single criteria seldom fully represent complex constructs, e.g., idea quality, they ought to measure such that additional in selecting a given criteria the research may introduce selection biases [67]. However, a single item may have sufficient predictive validity for any given dimension [5].

Additionally, process theories of survey response indicate that the presentation of ideas in an idea rating task – including the rating scale – affects the behavior of respondents [74]. Respondents act as cooperative communicators: they will endeavor to make sense of the questions by scrutinizing all information including formal features of the rating scale such as the numeric values or the graphical layout [68]. If respondents are unsure about what is being asked and face questions with no 'right' answer, such as rating idea quality, this behavior is intensified. Thus, the individual criteria of multi-criteria rating scales are strong cues for the rating of idea quality [12]. Theories on creativity indicate that these cues help respondents to judge idea quality more accurately. From research on brainstorming it is well understood that cognitive stimuli improve creative performance of individuals as they activate knowledge structures that have not been taken into account before [25, 36]. In this step of cognitive combination, information cues trigger an evaluation process in which the investigated creative concepts, e.g., innovation ideas, are constantly explored, interpreted, and assessed in order to apply them to a given context [19, 25]. In this regard, multi-criteria scales may activate associated knowledge structures that work as analogies to activate remote concepts in a user's cognitive web that help to develop a broader and more holistic understanding of the ideas to be evaluated.

Finally, research on multi-criteria decision making indicates that multiple criteria help users to develop shared mental problem representations as intended by the host of the online innovation community [51]. In this regard, multi-criteria scales may provide decisional guidance in terms of how the community host wants the rating scale users to think about idea quality. Moreover, additional rating criteria may break down the evaluation of idea quality into less complex sub tasks that address single aspects of idea quality [45]. A multi-criteria rating scale may better support the process of judging idea quality as different aspects of an idea are judged separately and mapped on different categories of the rating scale instead of integrating all aspects of the judgment in a single measure. This lowers the cognitive load of the idea rating and may thus improve decision quality of rating scale users. Thus, we assume:

H1: The rating scale used influences the decision quality of its user such that users of 'multi-criteria scales' have a higher decision quality than users of 'single criteria scales.'

Customer-generated new product ideas are creative products, which combine existing elements in a novel manner and must satisfy pre-existing criteria such as a fit with a firm's strategy and the needs of its customers. The ideas are the result of a non-deterministic creative process and yield semantic information that overlaps the information in the initial knowledge [43]. However, these ideas are often not very specific and show a rather low degree of elaboration and maturity resulting in vague and blurry descriptions [6].

In decision and management research, the direct relationship between such uncertain environmental conditions and decision quality has been well established. More accurate, understandable, and more comprehensive information enables decision makers to perform better decisions [71]. Thus, more elaborate ideas with comprehensive descriptions should be easier to evaluate than ideas that are very short and lack background information that would be needed for assessing idea quality. For rating scale users, the evaluation of less fully elaborated ideas may induce a high need for cognition and sense-making in order to derive an accurate evaluation of the idea quality. In these conditions, a multi-criteria scale might be more appropriate as it may provide decisional guidance for rating scale users leading to the integration of aspects of the ideas that are not mentioned in the idea description itself.

Prior research has well established the direct relationship between the quality of information used by a decision maker and the resulting decision quality [57]. Furthermore, existing research suggests that human decisions are mostly based on accessibility of information [57, 64] which suggests that most rating scale users take only these information into account that are present in the idea description. Thus, if there is less information used by the decision maker, such as in the case of less fully elaborated ideas, decision quality is also lower. Using a multi-attribute rating scale prompts users to answer questions about specific aspects of the idea quality construct such as its novelty or value and thus forces them to access information that may not be present in the idea description. Thus, we expect the multi-attribute rating scale to perform better in situations of low idea evaluation as they engage users to access additional information (e.g., from their own experience). On the flipside, more elaborated ideas tend to deliver more detailed information, e.g., regarding their novelty or feasibility, and thus already integrate these information into the idea description. For instance, more elaborate ideas are more likely to discuss its novelty compared to alternative solutions. Hence, there is less need for the decision maker to access additional information beyond what is already present in the idea description.

Thus, a single-criteria rating scale may be sufficient for the assessment of idea quality of well elaborated ideas. Summing up, we assume:

H2: Idea elaboration moderates the relationship between the rating scale and the decision quality of its users such that the gain in decision quality of 'multi-criteria scales' over 'single-criteria scales' will be lower for well elaborated ideas and higher for less elaborated ideas.

2.2. Perceptional Aspects of Rating Scales

Besides such functional aspects, the perception of rating scales has to be considered for creating effective rating scale designs. Ideally, rating scales should not only help community operators to filter out the most attractive ideas but also create favorable and enjoyable user experiences. Positive user experiences can be considered a central antecedent of future rating scale use and may also have effects on the perception of the entire website. Thus, appropriate rating scale design may help ensure a sufficient number of ratings for creating effective rankings. Besides, positive user experiences favor goals of an online innovation community's operator that may reach beyond effective rating and improving their innovative strength such as increasing customer loyalty, improving brand image, or recruiting new employees [28, 60].

Attitudes are frequently considered as surrogates for measuring the success of information systems, as they are directly rooted in the usage experience of users [29, 30, 33] and positively influence the usage of various information systems [42, 48, 76]. Attitudes are internal evaluations of a person towards a specific object and consist of affective, behavioral, and cognitive components [55, 70].

Such experience-based success measures are even more important in hedonic information systems like online innovation communities whose usage is based on the free will of users. Generally, research distinguishes utilitarian and hedonic information systems. For utilitarian systems, there is an external cause for usage such as improved efficiency. This generally applies to working situations where individuals have to use information systems as part of their daily work. Thus, technology adoption research proposes perceived usefulness and perceived ease of use to be central predictors of usage. In contrast, hedonic information systems aim to provide self-fulfilling value to users [35]. In this regard, behavioral intentions to use an information system emerge as plans for avoiding undesirable outcomes, thus increasing or maintaining positive outcomes based on the feelings that have been associated with using the information system in the past [46]. As people derive intrinsic value from enjoyable experiences, they try to maintain or re-experience such states of pleasure [16]. In this regard, positive user experiences are a pivotal driver for participation in creative activities, such as evaluating ideas with a given rating scale, and using hedonic information systems in general [35].

In this regard, we argue that multi-criteria scales create a more favorable attitude towards the website than single-criteria scales do and that this effect is mediated by attitude towards the rating scale. Mediating effects are defined as variables that explain the relationship between a predictor and an outcome in terms of 'how' the predictor is influencing the outcome [4, 59]. This mediation implies a direct positive effect of the rating scale on users' attitudes towards that rating scale, which we expect to occur due to two reasons.

Firstly, creativity research suggests that creative tasks like idea evaluation have to be considered as intrinsically enjoyable in order for participants to create a compelling usage experience as a consequence of using a rating scale [1]. In this regard, flow theory suggests that a positive usage experience of using information systems occurs when a given task's complexity is met by the individual's ability to solve it successfully [48]. This is true when the task is neither too easy nor too complex. If the task is considered to be too easy, the emotional consequence of solving the task will be boredom [44]. If users perceive the task to be too complex, they will be frustrated and dissatisfied, as they cannot cope with their expectation of getting the task done. Evidence from the neuropsychological literature suggests that cognitive judgments are generally accompanied by emotional ones [82]. This form of affective experience that we call 'feeling' accompanies all cognitive decisions that are formed through conscious thought. Thus, all judgments of objective properties, such as idea quality, are influenced by affective reactions [82]. Single-criteria rating scales force respondents to integrate all their cognitions of an idea into a single decision. Because idea quality and the emotions that arise during the decision-making process may be ambiguous, respondents may fail to integrate all affective and cognitive facets of their judgment into a single rating. Discrepancies between affective and cognitive evaluation of ideas may lead to high decisional stress and dissatisfaction [40]. Thus, users of a single-criteria rating scale are more likely to perceive a mismatch between the cognitive judgment of idea quality, emotions accompanying the rating, and their actual rating behavior leading to unfavorable attitude towards the rating scale than users of a multi-criteria rating scale.

Secondly, multiple criteria lead to an increase in interactivity. Interactivity can be defined by the possibilities of creating interactive content or messages on the website [38, 81] or interaction possibilities in general [61, 81]. Following this definition, multi-criteria rating scales can be considered as more interactive than single-criteria rating scales as they provide additional possibilities to rate the quality of the idea and express individual opinions. Increasing levels of interactivity alter the relationships between users and websites stimulating users and enriching interaction [38, 73]. Thus, more interactive websites have been positively associated with user satisfaction, enjoyment, motivation, and acceptance that are all pivotal determinants of positive user attitudes in online environments [81]. Thus, we assume:

H3: The rating scale used influences the attitude towards the rating scale of its user such that users of 'multi-criteria scales' have a more favorable attitude than users of 'single-criteria scales.'

Research from the fields of marketing and consumer behavior suggests that the formation of attitudes towards a specific object is highly determined by already existing attitudes [8, 42]. This holds true in online environments as well, where, e.g., the attitude towards the websites of traditional brick-and-mortar-retailers is highly influenced by attitudes formed offline [79]. The direct formation of attitudes towards a specific object is a high cognitive effort for individuals [58]. Thus, attitudes are often formed by attitudes of closely related objects. This attitude transfer is based on simple and intuitive interferences from cues that can be easily processed by individuals and stem from direct interaction [54]. As individuals seek to minimize their cognitive effort in attitude formation [23], the attitude towards the rating scale that is formed by the direct usage of the rating scale is likely to influence the attitude towards the website as well. We assume:

H4: The effect of the rating scales on a user's attitude towards the website is mediated by the user's attitude towards the rating scale.

Consolidating all four hypotheses the following research model emerges (Figure 1).

------------- Figure 1 about here -------------

3. Research Design

The study was designed as a between subject web experiment. The experimental factor had two levels: a single-criteria rating scale and multi-criteria rating scale. This resulted in two rating scale treatments between which rating scale users were randomly assigned. Within each group, we first collected system data of users' actual rating behavior. After completing the rating task, we collected perceptually anchored survey data from the participants. We used an independent expert rating of idea quality as a baseline for comparison to assess decision quality of rating scale users. The expert rating was collected before the experiment. Finally, we employed a bootstrap-based simulation to test the sensitivity of rating scales regarding the number of aggregated idea ratings and make additional assertions regarding the design of rating scales. The web experiment used a website developed by the authors through which users could submit their ratings (see Figure 2 and 3).

------------- Figure 2 and 3 about here -------------

The order of ideas on the website was randomized for each rating scale user so that all of them evaluated the ideas in a different order to avoid position bias. The random ordering was also used to make it more difficult for users to collaborate on the rating task, as the goal was to collect independent answers. Rating scale users performed the evaluation on their own computers via a web browser. Before starting the experiment, we tested whether all common browsers displayed the website in a similar way and no irregularities were discovered. As a web experiment closely reflects the actual usage scenarios of social participation mechanisms of websites, a high external validity of our results can be assured. Users rated the ideas in their natural environment and could allocate as much time as necessary to complete the task. The internal validity of results was assessed by analyzing the website's log files. This allowed us to identify users who had an improbable response behavior such as performing all ratings in less than five minutes. We also investigated the log files to look for indications that participants did not perform the rating task independently. We found no indication that this was the case. The website provided immediate visual feedback to the successful rating (i.e., the respective button was highlighted) which made it convenient for users to navigate through the system to identify ideas that had not yet been rated. Users could rate each idea only once. To study the effect of the rating scales and avoid confounding effects it is important that user ratings are independent [24]. Consequently, the rating information provided by other rating scale users was not visible. This is expected to increase rating accuracy [52]. The following sections explain the experimental design in more detail.

3.1. Idea Sample

The ideas to be evaluated in this experiment were taken from a real-world idea competition of a software company that was conducted in summer 2008 whose ideas were evaluated by the expert panel. The goal of the idea competition was to collect ideas how the main software product of the company, an enterprise resource planning system, could be improved. Both incremental and radical ideas where welcomed. The idea competition was targeted at users of the ERP system, as some degree of knowledge of the system was required. As a result, 57 ideas were submitted to the competition. All ideas contained only text and no images or other media content. Among these ideas, idea quality was normally distributed. The ideas varied in length between a half and a full type-written page. We drew a stratified sample of 24 ideas total of high, medium, and low quality based on the independent expert assessment with eight ideas drawn from each idea quality group. The sample size was considered sufficient as 20 to 30 ideas are generally used to measure the variance of creativity ratings in creativity research [9, 66].

3.2. Participants

Participants in topic-related online innovation communities are the target population of our experiment. Prior research has shown that people engaged in such communities are predominantly male, young, and well-educated [27, 41].

231 rating scale users participated in the experiment; 12 were excluded as they did not rate all ideas, did not fill out the questionnaire completely, or provided their ratings in less than five minutes. Our sample population consisted of undergraduate and graduate students from four information systems courses and research assistants in the department of information systems at a large German university. We considered information system (IS) students and research assistants to be appropriate rating scale users in this study because the experimental task required knowledge of enterprise software systems to judge idea quality. Furthermore, it can be argued that IS students are suitable as they represent actual users of online innovation communities. Moreover, Voich [77] found the values and beliefs of students to be representative of individuals in a variety of occupations. The mean age of the rating scale users was 22.16 years and 67% were male. The majority were undergraduate students (71%), while 23% possessed a bachelor degree and 6% a master degree.

3.3. Rating Scales

For each rating scale a different website was set up and made accessible under a unique URL. The rating scales comprised a single-criteria rating scale with five intervals and a multi-criteria rating scale. The single-criteria rating scale reflects an aggregated measure for idea quality. The multi-criteria rating scale consisted of four 5-point rating scales which reflect the key dimensions of idea quality: novelty ('How novel do you think this idea is?'), value ('What do you think is the value of this idea if implemented?'), feasibility ('How easy is it to implement this idea?'), and elaboration ('Is the idea well elaborated?'), [6].

3.4. Procedure

Rating scale users were randomly assigned to one of the two rating scales. We assigned them to one of the rating groups via a personalized email, which also included links to the experiment system and the online questionnaire. They completed the rating task during the following four weeks in November and December 2009. Participants were instructed that they had to complete the rating task individually and that collaboration was not allowed. The number of rating scale users in the single-criteria group was 103 and 116 in the multi-criteria group. Applying MANOVA we found no significant differences regarding age ($F= 0.24$; $p = 0.62$), gender ($F= 1.23$; $p = 0.27$), or university degree ($F= 0.03$; $p = 0.87$), so that randomization of rating scale users was successful.

4. Data Sources and Measures

The measures used in our analysis combine four data sources and collection methods in our study: (1) a web experiment reflecting users' idea evaluations, (2) a survey of rating scale users to gather perception and attitude, (3) an independent expert rating of idea quality, and (4) idea elaboration which was directly derived from the textual representation of the ideas being evaluated. Triangulation of these data sources allows detailed insights into the complex interaction of user behavior, satisfaction, and IT artifacts. Various researchers advocate the use of multiple methods and data sources to gain more robust results overcoming common method bias [18, 69].

4.1. Idea Evaluations from Rating Experiment

The 219 rating scale users performed 13608 ratings in total ((103 users * 1 rating + 116 users * 4 ratings) * 24 ideas). The median time required to rate the 24 ideas was 38 minutes and

22 seconds. It should be noted, that the time taken for submitting the ratings does not include the time spent reading through the idea.

4.2. Perception and Attitude Measures from Questionnaire

Attitude towards the rating scale as well as the website were measured using a post-experiment online questionnaire. For measuring attitude towards the rating scale and the website we adapted measures from Galletta et al. [29] and Geissler et al. [30] (cf. table 1).

------------- Table 1 about here -------------

The questionnaire was pretested with a sample of ten subjects that reflected the group of rating scale users and led to minor changes to the questionnaire. All items were measured with a 5-point Likert scale. Performing exploratory factor analysis with SPSS 19.0 we tested the dimensional structure of our attitude scales (cf. table 2). Except item ATW4 that had to be excluded from analysis, all items loaded unambiguously on two factors that can clearly be interpreted. We checked whether the data was appropriate for explanatory factor analysis by calculating the Measures of Sampling Adequacy (MSA) for the whole data structure as well as for individual items. As all MSA values were above 0.6, exploratory factor analysis was applicable and no items had to be eliminated. The reliability of the factors was checked using Cronbach's Alpha. Alpha should be higher than 0.7 for indicating an acceptable value for internal consistency [56]. With Alphas of at least 0.75 this criterion can be considered as met. Subsequently, we tested these factors applying confirmatory factor analysis using Amos 19.0 (cf. Table 2). Initially multivariate normality was confirmed, so that Maximum-Likelihood-Estimation could be applied. The two factors showed very high Composite Reliabilities and high values for the Average Variance Explained (AVE), so that convergent validity can be assumed (cf. Table 2). Values of 0.6 regarding the Composite Reliability and 0.5 for the AVE can be seen as minimum values for indicating a good measurement quality [3]. The discriminant validity of the factors was checked by using the Fornell-Larcker criterion which claims that the square root of one factor's AVE should be higher than its correlation with every other factor [26]. As the two square roots of both factors (ATR = 0.70; ATW = 0.72) exceeds the correlation between the two constructs (0.28), discriminant validity can be assumed. For all items, Individual Item Reliabilities were calculated that were all above the minimum threshold of 0.4 [3]. Overall, the scale's good reliabilities based on Cronbach Alpha can be confirmed.

------------- Table 2 about here -------------

Finally, we checked the global fit of our measurement model by conducting a Chi-Square ($\chi 2$)-test. The χ^2-test was significant ($p = 0.00$) and the χ^2 / df-ratio was 3.21, well below the upper threshold of 5.00, which indicates good fit [80]. Furthermore, global fit measures suggested excellent fit as well: GFI = 0.95 (Goodness of Fit Index; ≥ 0.9), AGFI = 0.88 (Adjusted Goodness of Fit Index; ≥ 0.9), NFI = 0.95 (Normed Fit Index; ≥ 0.95), CFI = 0.95 (Comparative Fit Index; ≥ 0.95), and SRMR = 0.05 (Standardized Root Mean Square Residual; ≤ 0.11). Thus, the scales were successfully validated using both exploratory and confirmatory factor analysis.

4.3. Expert Rating

To assess decision quality of rating scale users, the users' ratings were compared with the independent expert rating to assess their judgmental fit. All 57 ideas of the idea competition were evaluated by a qualified expert jury using the consensual assessment technique [1] which has been used to assess the quality of customer-generated new product ideas [53]. The jury consisted of seven referees with high expertise relevant to the field of the ideas. The complex construct of idea quality was operationalized with four dimensions and measured by 15 items ranging from one (lowest) to seven (highest). The rating tasks were distributed in randomized order on paper-based forms including the idea descriptions. Based on the 57 ideas, we conducted exploratory and confirmatory factor analysis in order to assess the validity and the reliability of the expert rating. Due to this analysis six items were removed. The Intra-Class-Correlations for remaining item are 0.6 or higher. We followed the procedure described in [6].

4.4. Idea Elaboration from Textual Idea Descriptions

We counted the number of characters of the textual representation of an idea as a measure of idea elaboration. The measure includes all whitespace (length of title plus description measured in characters). All ideas were represented only as text and did not contain other content such as images or videos. Consequently, the length of this textual representation captures the actual length of an idea and can thus be a good indicator for idea elaboration.

5. Analysis

5.1. Testing Functional Aspects of Rating Scales

Following the wisdom of crowds paradigm, the idea rankings that are produced from aggregating all ratings of all users are most important for companies operating open innovation communities. Thus, we analyzed the idea rankings for each rating scale when aggregating the individual user ratings for each idea and compared them to the aggregated expert rating for that idea in order to test H1. We performed (1) a Kendall-Tau correlation analysis and (2) an error measurement commonly used in time series analysis. Using correlation analysis with the expert ranking, the multi-criteria rating scale showed a significant concurrence with $r = 0.47$ ($p < 0.01$; Table 3). The ranking produced by the single-criteria rating scale did not correlate with the expert ranking significantly ($r = 0.02$). Also, the rankings produced by both rating scales show no correlations among each other ($r = 0.22$; $p < 0.001$). In addition to the correlation analysis, we used the mean absolute percentage error (MAPE), a dimensionless metric, as a measure for the deviation of the results of each rating scale from the independent expert rating [2]. While the correlation analysis allows the assessment of overall consent of the experiment ratings with the independent expert rating, the analysis using MAPE allows a relative comparison between the two treatment groups. MAPE is the most widely used measure for evaluating the accuracy of forecasts in time series analysis and offers good validity [72]. We used the placement numbers of the aggregated ranking (by mean) and the sorted idea ranking as the fitted time series. Our application of MAPE is defined by the following formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{actual\ ranking\ of\ idea\ (expert\ rating)_t - forecast\ ranking\ of\ idea\ (experiment\ rating)_t}{actual\ ranking\ of\ idea_t}\right|$$

This measure imposes a higher penalty on errors in the ordering of good ideas than on errors in the ordering of bad ideas and corresponds to the economic aim of correctly identifying the good ideas. The smaller the MAPE, the smaller the deviation from the expert results. In case of a perfect fit, MAPE is zero but in regard to its upper limit there is no restriction. The expected value of MAPE for a random ranking of the 24 ideas is 1.45. The multi-criteria rating scale resulted in a MAPE of 1.02; single-criteria rating scale was 1.43 (Table 3).

A key issue when collecting information from multiple informants is how the data is aggregated. Marketing research has shown that applying other methods than averaging informants' responses can improve accuracy. Thus, we checked if the aggregated idea ranking could be improved by varying weights in the aggregation of user ratings. We implemented an aggregation method based on response data-based weighted means [75] that reduces the impact of systematic measurement errors. In this weighted aggregation method, ratings of agreeing users are weighted more heavily than ratings of users who lack agreement and deviate from the mean. Using the default configuration of a uniform adjustment variable $\alpha = 1$ [75], the MAPE of the single-criteria rating scale increased slightly to 1.50 (i.e., worsened), and the MAPE of the multi-criteria rating scale decreased to 0.95 (Table). Summarizing all findings, H1 can be accepted.

------------- Table 3 about here -------------

As our data has in general a hierarchical structure (each user has performed 24 ratings), we initially tested as whether an analysis of this hierarchically structured data required dedicated analysis methods such as Hierarchical Linear Modeling. However, this is not the case as only very low Intra-Class-Correlations (< 0.01) were found. Thus, the analysis of the moderating effect of idea elaboration between rating scale and decision quality (H2) was performed on the level of the individual ratings only. Single idea ratings serve as unit of analysis.

For the operationalization of decision quality, we applied a procedure from creativity research in which judgmental accuracy of laypersons is usually determined by assessing the concurrent validity of their judgments with those of an expert jury, e.g., by counting 'good ideas' or 'bad ideas' that have been identified correctly by the non-experts [66]. We adapted this approach to measure decision quality of user ratings. Current research about customer-generated new product ideas shows that about 30% of ideas are of high quality [6]. Thus, we defined the eight ideas with the highest quality according to experts as 'good ideas' and the eight ideas with the lowest quality according to the experts as 'bad ideas,' respectively. In a similar vein, we repeated this classification for each user (i.e., the eight ideas that a given user rated best ('worst') were considered as 'good' ('bad')). As a second step, we awarded decision quality scores for each idea rating. An idea was correctly classified when a 'good idea' of a given user has also been considered as 'good' by the experts. Additionally, we considered ideas as correctly classified if a user's 'bad ideas' were among the eight ideas with lowest quality according to the experts. An idea was considered as wrongly classified when a 'good' user idea was classified as 'bad' by the experts and vice versa. Correctly classified ideas were awarded a value of 1 and wrongly classified ideas a value of 0. Ideas that were neither correctly, nor wrongly classified were deleted from the sample. As our dependent variable, decision quality, reflected a dummy variable we applied logistic regression for assessing H2 [34]. Our experimental condition, rating scale, was also reflected as a dummy variable in which the multi-criteria scale served as reference group. Idea elaboration was measured in characters. We found a statistically significant negative moderation ($p < 0.01$) (cf. Table 4) and we plotted the estimated means for visually verifying this result (Figure 4). This indicates that the potential decision support a multi-criteria rating scale may endow is limited by an idea's elaboration. Thus, H2 can be supported.

------------- Table 4 about here -------------

------------- Figure 4 about here -------------

5.2. Testing Perceptional Aspects of Rating Scales

For testing the direct effect of single- and multi-criteria scales on attitude towards the website (H3) and the mediation effect of attitude towards the rating scale between the rating scales and attitude towards the website (H4), we applied the procedure of Preacher and Hayes [59] which uses a direct bootstrapping-based (N=500) significance test for mediation. For this analysis, users were our unit of analysis. Mediation can be detected with multiple OLS regressions and is generally assumed if the strength of the relationship between a predictor (rating scale) and an outcome variable (attitudes towards the website) diminishes in case a third mediator variable (attitudes towards the rating scale) that is caused by the predictor and influences the outcome is entered into the regression. This implies a direct effect of the rating scales on attitude towards the rating scale as well, which is significant in our case so that H3 is supported (cf. Table 5, step 1). Moreover, the regression coefficient of the direct effect of the rating scale on attitude towards the website is declining from 0.11 to -0.03 in case attitude towards the rating scale is entered into the regression equation (cf. Table 5, step 2 and 3). This decline is significant with $p < 0.05$ thus supporting H4.

------------- Table 5 about here -------------

5.3. Sensitivity Analysis and Simulation

While the research model developed and tested above investigates the effects of single- and multi-criteria rating scales, a central issue in the measurement of user responses is to determine the number of responses necessary to arrive at stable results [75]. This is particularly so in online innovation communities where participation fluctuates and the host of the community is predominantly interested in the aggregated ranking of the user ratings. The ability to elicit high participation and control rating scale users so that they make valuable contributions is thus an important challenge in the design of social interaction systems [60]. Following research on aggregation of individual opinions [52, 75], a smaller amount of available idea ratings should decrease the quality of the aggregated idea rating because less, and less diverse, information is available. As a consequence, individual decision errors have a stronger weight [67]. Thus, we use our data to simulate the number of ratings necessary to construct stable aggregated idea rankings.

In order to assess the impact of the number of ratings available for aggregation on the stability of the resulting ranking, we performed a Monte Carlo approximation to the bootstrap estimate to determine how many user ratings (per rating object) are necessary to arrive at stable overall rankings. Monte Carlo simulations are a class of computational algorithms that rely on repeated random sampling to compute their results [65]. Bootstrapping relies on the logic of re-sampling from the original dataset to approximate the distribution of parameters [47]. Different Monte Carlo and bootstrap simulations have been reported in IS research [e.g., 11, 32]. We used this re-sampling-based simulation to approximate the sensitivity of user ratings by following a general process common to most Monte Carlo approximations of bootstrap estimators in which we re-sample from the users' original ratings to estimate aggregate rankings of the pool of 24 ideas [10].

In our simulation approach we randomly drew ratings (with replacement) from the original dataset of ratings, aggregated these ratings (using means), and created a ranked list of the 24 ideas. We then calculated the MAPE by comparing the newly created ranking to that of the independent experts (the same we did in the analysis above). We repeated these steps drawing N = 0, 1, 2, 3, … 2500 individual ratings (approximately the size of the original dataset). For the initial, small re-samples that did not contain a rating for every idea, we randomly ranked ideas without a rating. Consequently, the simulation starts with the expected MAPE for a random ranking of the 24 items (1.45). We randomly ranked ideas where the exact rank order could not be determined due to a draw. This simulation was performed for both rating scales. Figure 5 shows the Monte Carlo approximation of 100 simulation runs for each of the two scales as well as the MAPE of a random ranking at the 1.45 mark. While the single-criteria rating scale performs only slightly better than random (MAPE of 1.38 after 2500 randomly drawn ratings) the multi-criteria rating scale, starting with the random MAPE of 1.45, drops sharply with each additional rating to an accuracy substantially better than random and then slowly converges towards a final value of 0.98. The curve drops steepest in the beginning and levels off after around 20 ratings per idea. In particular, there is only a performance increase of 7% to be gained by moving from 20 ratings per idea to 100 ratings per idea.

From the regression analysis we know that idea elaboration has significant influence on the gain in decision quality of multi-criteria scales over single-criteria scales. We test this influence in the simulation by combining user ratings from both rating scales. Specifically, in the bootstrap-based simulation, we randomly draw user ratings from the single-criteria treatment for highly elaborated ideas and user ratings from the multi-criteria rating scale for less elaborate ideas. The simulation thus shows how the moderating effect of idea elaboration can be exploited by combining the single-criteria rating scale and the multi-criteria rating scale. The combined approach results in a MAPE of 0.85, which is a 63% performance improvement over only the single-criteria rating scale and 16% over only the multi-criteria rating scale.

5.4. Robustness of Analysis

In our analysis of the main condition of interest, decision quality, we used three different analysis methods: correlation analysis with the expert ranking, error measurement using MAPE, and a simulation-based approach. The results of all analysis agree and indicate that the multi-criteria rating scale performs significantly better (highest correlation with expert ranking, and lowest MAPE). To test the robustness of our individual user analysis, we used a five and eight idea cut-off criteria, which lead to almost identical results. The additional analysis using MAPE allows for a convenient direct performance comparison of the aggregated results across the rating scales. In summary, the individual user scores agree with the aggregated results (both correlation and MAPE).

In order to support the validity and robustness of our results, we performed an additional analysis on the aggregated idea level using response data-based weighted mean aggregation. Here, the MAPE of the single-criteria rating scale worsened slightly, and the MAPE of the multi-criteria rating scale improved (cf. Table 5). This indicates that the measures do not contain systematic measurement errors, which would have been eliminated using the weighted means aggregation method. This supports the robustness of the results as they are not dependent on the simple unweighted mean-based aggregation and strengthens our position regarding the improved performance of the multi-criteria rating scale. Finally, in addition to using multiple analysis methods, we also used a multi-method approach to collect data comprising system-captured user ratings, perceptually anchored, self-reported user data, and an independent expert rating. This multi-method approach exhibits a low susceptibility to common method variance and provides richer data for analysis and more reliable results [69].

------------ Figure 5 about here -------------

6. Summary and Contribution

Using system-captured experiment data, perceptually anchored questionnaire data, and an independent expert evaluation of idea quality, the proposed theoretical model was tested for the functional and perceptional effects of single- and multi-criteria rating scales. We found that the multi-criteria rating scale leads to higher decision quality in comparison to the single-criteria rating scale, supporting H1. It was expected that idea elaboration would have a moderating effect on the relationship between the rating scale and decision quality. This moderating effect of idea elaboration was supported (H2). We also tested for a mediating effect of attitude towards the rating scale between the rating scales and attitude towards the website. This was also supported (H3 and H4) indicating that users' perceive the multi-criteria rating scale more favorable than the single-criteria scale. Finally, using a bootstrap-based simulation we first showed that an average of around 20 ratings per idea leads to stable rankings. Adding additional ratings increases the accuracy only slightly: a performance increase of only 7% can be gained by moving from 20 ratings per idea to 100 ratings per idea. Second, our simulation shows how a 16% performance improvement in decision quality could be achieved by exploiting the moderating effect of idea elaboration through combining single-criteria ratings for long ideas and multi-criteria ratings for short ideas.

The use of system-captured experiment data, questionnaire data, and independent expert ratings offers a fuller appreciation of the phenomena under investigation that would not have been possible using a single data source only. The use of multiple data sources was further extended using multiple levels of analysis and analysis methods. Overall, there is mutual support between the methods of analysis and data sources. Simulation results in particular add to our knowledge as to (1) how sensible aggregate measures of a given rating scale are towards the number of available idea ratings, and (2) the potential performance improvement of combining the two rating scales compared in this study based on the moderating effect of idea elaboration.

6.1. Theoretical Implications

With the surge of social interaction and user-generated content on the Internet, website design with appropriate application of rating scales is an important topic for research. Understanding underlying mechanisms is key to systematic design rating scales. These scales have direct effects on both the effectiveness of the resulting user ratings and the perception of these rating scales as a predictor of future website use. Consequently, both functional and perceptional aspects have been investigated in this research.

The present study adds to the existing work on website design with a spotlight on rating scales. While different rating scales have been examined, in particular in marketing research [e.g., 13, 56], the effectiveness of these mechanisms in an online context has not been well determined yet [7, 84]. This research, therefore, contributes to the discussion of co-creation and underlying mechanisms to leveraging the potential of user-generated content as to how one specific element of website design – rating scales – impact outcome and perception. While several important elements of website design such as the use of human images [18] or product presentation formats [42] have been studied, rating mechanisms which have become a key concept in many current websites have not yet been studied in detail.

Specifically, this research contributes to our understanding of the interaction of the technology being used (i.e., which rating scale), and attributes of the rating object on two central outcome measures: the effectiveness of the rating in terms of decision quality of its user and the perception of the scale by the user as a predictor of future use. While our finding of superior effectiveness of the multi-criteria rating are well reflected in existing scale literature [12, 74] our analysis employs a broader perspective taking attributes of the rating object (idea elaboration) into account. Furthermore, given that importance of website usability and the consequential tendency of web designers to employ the simplest, most user friendly rating scales, our study puts a rating scale's effectiveness into this broader perspective. Thus, earlier general findings are now applied in the realm of online rating scales in which additional considerations play an important role in the design of the overall interaction system.

Our results have general application in contexts in which only a small fraction of a larger number of ideas is valuable such as brainstorming sessions [31], communities [22], or contests [6, 50].

6.2. Practical Implications

The design of rating scales on websites is critical as it influences both rating outcomes as well as users' attitudes and thus their intention to use a website. Our research suggests that for hosts of online innovation communities significantly improved results can be achieved by combining multiple rating scales. A simple measure such as text length can be used to implement a dynamic system that would present users with different rating scales depending on the degree of elaboration of the idea. This allows exploiting the moderating effect of idea elaboration and thus improves the effectiveness of user ratings. Furthermore, our simulation shows that with an average of as little as 20 ratings per rating object a stable ranking can be achieved. This is of important practical relevance as until now it is unknown when stable rankings can be constructed from website ratings.

However, the practical value of our results depends on the costs of idea evaluation in regard to the potential of the ideas, the type I error and more importantly the type II error that is associated with the idea evaluation. In innovation, the costs associated with wrongly classifying a bad idea as good (type I error) can be significantly different from the costs of wrongly classifying a good idea as bad (type II error). While implementing ideas in the former case simply reflects a misallocation of financial resources, the latter case may reflect a lost opportunity, which can be fatal to the focal company. The risk of occurrence and the consequences of misclassification errors generally rise with the concentration of idea quality on a small number of very good ideas. However, existing research shows that the decision quality of laymen rise in such conditions of high variance of the rating objects' quality [9]. Moreover, the best concepts in online innovation communities are generally crystallization points of intense discussions of community members [22]. Thus, type II errors can be minimized in practice when the focal company also takes these qualitative discussions into account. On the flipside, the negative consequences of misclassifications become less severe the higher the aggregated costs of idea evaluation are for all ideas. As several thousand ideas are quickly contributed to successful online innovation communities, and experts are a scarce resource, these costs are highly relevant.

There is a tradeoff between offering a simpler but in some cases less predictive rating scales and more complex rating scales that are able to collect richer information but might put a higher burden on its users thus possibly reducing future use. While this is true in the general case and would possibly sway businesses against using more complex rating scales, our analysis finds that the more complex rating scale created more favorable and enjoyable user experiences, a prime antecedent of future use. Consequently, we argue that the multi-criteria rating scale can not only leads to higher decision quality but can also lead to more favorable and enjoyable user experiences.

6.3. Limitations and Directions for Future Research

Some general shortcomings resulting from conducting an experiment apply to our research. Users were not allowed to choose the ideas to be rated. However, this should not lead to a significant distortion as both rating scales offer a neutral rating option. Furthermore, given the design of some incentive schemes that are based on overall user activity, a user rating a majority of ideas in a system does not constitute an unlikely setting. Our experimental context did not allow us to measure differences in the level of user participation between the two rating scales. It could be possible that the more detailed multi-criteria scale would lead to lower levels of user participation as it puts a higher burden on its users. However, given that the multi-criteria scale was perceived more positively this might not be the case and high enjoyment might even lead to higher levels of user participation. Future research should investigate levels of participation between different rating scales. A second limitation results from our experimental design in which users could not see other users' ratings. While this was a deliberate decision based on results of prior work on the social influence of users [24, 52], we acknowledge that this decision leads to a slight deviation from real-world websites in which user rating would generally be visible. However, we found it more important to focus on the main condition of interest – the effects of the design of a particular website feature – without introducing additional confounding effects such as social influence and information cascades. Future research could extend the model tested in this research by explicitly adding experimental conditions to study the effect of social influence and information cascades. Finally, as the experiment was conducted as a web experiment there is the possibility of some bias as users might have collaborated on the rating task. However, given the clear instructions stating that the rating task had to be completed independently, given the randomized order in which ideas were displayed to each user, and log file analysis, we believe potential bias is at most marginal.

7. Conclusion

While rating scales are almost omnipresent in social participation and co-creation websites, they serve a particular purpose in online innovation communities that aim to use them as a filter mechanism to separate good from bad ideas. Thus, while designing those rating scales both functional and perceptional aspects need to be taken into consideration and a balance has to be struck between designs that work well and design that result in high users enjoyment and participation. Our work contributes to the larger stream of research investigating the design of co-creation mechanisms and websites in general. We hope that other researchers join our efforts and collectively we can deepen our understanding of the various elements and underlying mechanisms that govern consumer co-creation.

References

1. Amabile, T.M. *Creativity in context. Update to social psychology of creativity*. Oxford: Westview Press, 1996.
2. Armstrong, J.S., and F. Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons *International Journal of Forecasting,* 8, 1 (1992), 69-80.
3. Bagozzi, R.P., and Y. Yi. On the Evaluation of Structural Equatation Models. *Journal of the Academy of Marketing Sciences,* 16, 1 (1988), 74-94.
4. Baron, R.M., and D.A. Kenny. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology,* 51, 6 (1986), 1173-1182.
5. Bergkvist, L., and J.R. Rossiter. The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research,* 44, 2 (2007), 175-184.
6. Blohm, I.; U. Bretschneider; J.M. Leimeister; and H. Krcmar. Does collaboration among participants lead to better ideas in IT-based idea competitions? An empirical investigation. *International Journal of Networking and Virtual Organizations,* 9, 2 (2011), 106-122.
7. Bonabeau, E. Decisions 2.0: The power of collective intelligence. *MIT Sloan Management Review,* 50, 2 (2009), 45-52.
8. Bruner II, G.C., and A. Kumar. Web Commercials and Advertising Hierarchy-of-Effects. *Journal of Advertising Research,* 40, 1/2 (2000), 35-42.

9. Caroff, X., and M. Besançon. Variability of creativity judgments. *Learning and Individual Differences,* 18, 4 (2008), 367-371.

10. Chernick, M. *Bootstrap Methods: A Guide for Practitioners and Reseachers*. Hoboken, NJ, USA: Wiley, 2008.

11. Chin, W.W.; B.L. Marcolin; and P.R. Newsted. A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research,* 14, 2 (2003), 189-217.

12. Christian, L.M.; D.A. Dillman; and J.D. Smyth. Helping Respondents Get it Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys. *Public Opinion Quarterly,* 71, 1 (2007), 113-125.

13. Churchill, G.A. A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research,* 16, 1 (1979), 64-73.

14. Clemons, E.K.; G. Gao; and L.M. Hitt. When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems,* 23, 2 (2006), 149-171.

15. Cohen, W.M., and D.A. Levinthal. Absorptive Capacity: A New Perspective On Learning And Innovation. *Administrative Science Quarterly,* 35, 1 (1990), 128-152.

16. Csikszentmihalyi, M. *Creativity: Flow and the Psychology of Discovery and Invention*. New York, NY: HarperPerennial, 2002.

17. Cui, G.; H.-K. Lui; and X. Guo. The Effect of Online Consumer Reviews on New Product Sales. *International Journal of Electronic Commerce,* 17, 1 (2012), 39-57.

18. Cyr, D.; M. Head; H. Larios; and B. Pan. Exploring Human Images in Website Designs: A Multi-Method Approach. *MIS Quarterly,* 33, 3 (2009), 530-566.

19. Dailey, L., and M.D. Mumford. Evaluative aspects of creative thought: Errors in appraising the implications of new ideas. *Creativity Research Journal,* 18, 3 (2006), 367-384.

20. Dellarocas, C.; G. Gao; and R. Narayan. Are consumers more likely to contribute online reviews for hit or niche products? *Journal of Management Information Systems,* 27, 2 (2010), 127-157.

21. Dennis, A.R., and B.H. Wixom. Investigating the Moderators of the Group Support Systems Use with Meta-Analysis. *Journal of Management Information Systems,* 18, 3 (2001), 235-257.

22. Di Gangi, P.M., and M.M. Wasko. Steal my idea! Organizational adoption of user innovations from a user innovation community: A case study of Dell IdeaStorm. *Decision Support Systems,* 48, 1 (2009), 303-312.

23. Eagly, A., and S. Chaiken. *The psychology of attitudes*. Fort Worth, TX, USA: Harcourt Brace Jovanovich College Publishers, 1993.

24. Easley, D., and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, MA, USA: Cambridge University Press, 2010.

25. Finke, R.A.; T.B. Ward; and S.M. Smith. *Creative cognition. Theory, research and applications*. Cambrigde, MA, USA: MIT Press, 1996.

26. Fornell, C., and D.F. Larcker. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research,* 18, 2 (1981), 39-50.

27. Franke, N., and S. Shah. How communities support innovative activities: An exploration of assistance and sharing among end-users. *Research Policy,* 32, 1 (2003), 157-178.

28. Fuchs, C., and M. Schreier. Customer Empowerment in New Product Development. *Journal of Product Innovation Management,* 28, 1 (2011), 17-32.

29. Galletta, D.F.; R. Henry; S. McCoy; and P. Polak. Web Site Delays: How Tolerant are Users? *Journal of the Association for Information Systems,* 5, 1 (2004), 1-24.

30. Geissler, G.L.; G.M. Zinkhan; and R.T. Watson. The Influence of Home Page Complexity on Consumer Attention, Attitudes and Purchase Intent. *Journal of Advertising,* 35, 2 (2006), 69-80.

31. Girotra, K.; C. Terwiesch; and K.T. Ulrich. Idea Generation and the Quality of the Best Idea. *Management Science,* 56, 4 (2010), 591-605.

32. Goodhue, D.; W. Lewis; and R. Thompson. Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators. *Information Systems Research,* 18, 2 (2007), 211-227.

33. Goodhue, D., and R. Thompson. Task-technology fit and individual performance. *MIS Quarterly,* 19, 2 (1995), 213-236.

34. Hayes, A.F., and J. Matthes. Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods,* 41, 3 (2009), 924-936.

35. Heijden, H.v.d. User Acceptance of Hedonic Information Systems. *MIS Quarterly,* 28, 4 (2004), 695-704.

36. Hender, J.M.; D.L. Dean; T.L. Rodgers; and J.F. Nunamaker. An Examination of the Impact of Stimuli Type and GSS Structure on Creativity: Brainstorming Versus Non-Brainstorming Techniques in a GSS Environment. *Journal of Management Information Systems,* 18, 4 (2002), 59-85.

37. Hill, S., and N. Ready-Campbell. Expert Stock Picker: The Wisdom of (Experts in) Crowds. *International Journal of Electronic Commerce,* 15, 3 (2011), 73-102.

38. Hoffman, D.L., and T.P. Novak. Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. *Journal of Marketing,* 60, 3 (1996), 50-68.

39. Jacoby, J. Information Load and Decision Quality: Some Contested Issues. *Journal of Marketing Research,* 14, 4 (1977), 569-573.

40. Janis, I.L., and L. Mann. *Decision Making. A Psychological Analysis of Conflict, Choice, and Commitment*. New York, NY, USA: The Free Press, 1977.

41. Jeppesen, L.B., and L. Frederiksen. Why do users contribute to firm-hosted user communities? The case of computer-controlled music instruments. *Organization Science,* 17, 1 (2006), 45-63.

42. Jiang, Z., and I. Benbasat. Investigating the Influence of the Functional Mechanisms of Online Product Presentations. *Information Systems Research,* 18, 4 (2007), 454-470.

43. Johnson-Laird, P.N. *Human and machine thinking*. Hillsdale: Lawrence Erlbaum Associates, 1993.

44. Kamis, A.; M. Koufaris; and T. Stern. Using an attribute-based decision support system for user-customized products online: An experimental Investigation. *MIS Quarterly,* 32, 1 (2008), 158-177.

45. Keeney, R.L. *Value-focused thinking: A path to creative decision-making*. Cambrige, MA, USA: Harvard University Press, 1992.

46. Kim, H.W.; H.C. Chan; and Y.P. Chan. A Balanced Thinking-Feeling Model of Information Systems Continuance. *International Journal of Human-Computer Studies,* 65, 6 (2007), 511-525.
47. King, G.; M. Tomz; and J. Wittenberg. Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science,* 44, 2 (2000), 347-361.
48. Koufaris, M. Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior. *Information Systems Research,* 13, 2 (2002), 205-223.
49. Kwon, O., and Y. Sung. Shifting Selves and Product Reviews: How the Effects of Product Reviews Vary Depending on the Self-Views and Self-Regulatory Goals of Consumers. *International Journal of Electronic Commerce,* 17, 1 (2012), 59-91.
50. Leimeister, J.M.; M. Huber; U. Bretschneider; and H. Krcmar. Leveraging Crowdsourcing - Activation-Supporting Components for IT-based Idea Competitions. *Journal of Management Information Systems,* 26, 1 (2009), 197-224.
51. Limayem, M., and G. DeSanctis. Providing Decisional Guidance for Multicriteria Decision Making in Groups. *Information Systems Research,* 11, 4 (2000), 386.
52. Lorenz, J.; H. Rauhut; F. Schweitzer; and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of National Acadamey of Science,* 108, 22 (2011), 9020-9025.
53. Matthing, J.; P. Kristensson; A. Gustafsson; and A. Parasuraman. Developping succesful technology-based services: The issue of identifying and involving innovative users. *Journal of Services Marketing,* 20, 5 (2006), 288-297.
54. Meyers-Levy, J., and P. Malaviya. Consumers' Processing of Persuasive Advertisements: An Integrative Framework of Persuasion Theories. *Journal of Marketing,* 60, 1 (1999), 45-60.
55. Mitchell, A.A., and J.C. Olson. Are product attribute beliefs the only mediator of advertising effects on brand attitude? *Journal of Marketing Research,* 18, 3 (1981), 318-332.
56. Nunnally, J.C., and I.H. Bernstein. *Psychometric Theory*. New York, NY, USA: McGraw-Hill, 1994.
57. O'Reilly III, C.A. Variations in Decision Makers' Use of Information Sources: The Impact of Quality and Accessibility of Information. *Academy of Management Journal,* 25, 4 (1982), 756-771.

58. Petty, R.E., and J.T. Cacioppo. The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology,* 19, (1986), 123-162.
59. Preacher, K.J., and A.F. Hayes. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers,* 36, 4 (2004), 717-731.
60. Preece, J., and B. Shneiderman. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction,* 1, 1 (2009), 12-32.
61. Rafaeli, S. Interactivity: From new media to communication, In: *Sage Annual Review of Communication Research: Advancing Communication Science*, R.P. Hawkins, J.M. Wiemann, and S. Pingree, Editors. 1988.
62. Riedl, C.; I. Blohm; J.M. Leimeister; and H. Krcmar. Rating Scales for Collective Intelligence in Innovation Communities: Why Quick and Easy Decision Making Does Not Get it Right. in *International Conference on Information Systems (ICIS'10)*. 2010. St. Louis, MI, USA.
63. Riedl, C.; N. May; J. Finzen; S. Stathel; V. Kaufman; and H. Krcmar. An idea ontology for innovation management. *International Journal on Semantic Web and Information Systems,* 5, 4 (2009), 1-18.
64. Rieh, S.Y. Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology,* 53, 2 (2002), 145-161.
65. Rubinstein, R.Y., and D.P. Kroese. *Simulation and the Monte Carlo Method*. Hoboken, NJ, USA: Wiley, 2008.
66. Runco, M.A., and M. Basadur. Assessing Ideational and Evaluative Skills and Creative Styles and Attitudes. *Creativity and Innovation Management,* 2, 3 (1993), 166-173.
67. Rushton, J.P.; C.J. Brainerd; and M. Pressley. Behavioral Development and Construct Validity: The Principle of Aggregation. *Psychological Bulletin,* 94, 1 (1983), 18-38.
68. Schwarz, N. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Hillsdale, NY, USA: Erlbaum, 1996.
69. Sharma, R.; P. Yetton; and J. Crawford. Estimating the effect of common method variance: The method-method pair technique with an illustration from TAM research. *MIS Quarterly,* 33, 3 (2009), 473-490.

70. Solomon, M.; G. Bamossy; S. Askegaard; and M.K. Hogg. *Consumer Behavior. An European Perspective*. Harlow, UK: Prentice Hall International, 2006.
71. Streufert, S. Effects of information relevance on decision making in complex environments. *Memory & Cognition,* 1, 3 (1973), 224-228.
72. Tayman, J., and D.A. Swanson. On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review,* 18, 4 (1999), 299-322.
73. Teo, H.-H.; L.-B. Oh; C. Liu; and K.-K. Wei. An empirical study of the effects of interactivity on web user attitude. *International Journal of Human-Computer Studies,* 58, 3 (2003), 281-305.
74. Tourangeau, R.; L.J. Rips; and K. Rasinski. *The Psychology of Survey Response*. Cambridge, MA, USA: Cambridge University Press, 2000.
75. Van Bruggen, G.; G. Lilien; and M. Kacker. Informants in organizational marketing research: Why use multiple informants and how to aggregate responses *Journal of Marketing Research,* 39, 4 (2002), 469-478.
76. Venkatesh, V.; M.G. Morris; G.B. Davis; and F.D. Davis. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly,* 27, 3 (2003), 425-478.
77. Voich, D. *Comparative Empirical Analysis of Cultural Values and Perceptions of Political Economy Issues*. Westport, CT, USA: Praeger, 1995.
78. Von Hippel, E. *Democratizing innovation*. Cambridge, MA, USA: MIT Press, 2005.
79. Wang, S.; S.E. Beatty; and D.L. Mothersbaugh. Congruity's role in website attitude formation. *Journal of Business Research,* 62, 6 (2009), 609-615.
80. Wheaton, B.; B. Muthén; D.F. Alwin; and G.F. Summers. Assessing reliability and stability in panel models, In: *Sociological Methodology*, D.R. Heise, Editor. 1977, Jossey-Bass: San Francisco.
81. Wu, G. The Mediating Role of Perceived Interactivity in the Effect of Actual Interactivity of Attitude toward the Website. *Journal of Interactive Advertising,* 5, 2 (2005), 29-39.
82. Zajonc, R. Feeling and thinking: Preferences need no references. *American Psychologist,* 35, 2 (1980), 151-175.
83. Zheng, H.; D. Li; and W. Hou. Task Design, Motivation, and Participation in Crowdsourcing Contests. *International Journal of Electronic Commerce,* 15, 4 (2011), 57-88.

84. Zwass, V. Co-Creation: Toward a Taxonomy and an Integrated Research Perspective. *International Journal of Electronic Commerce,* 15, 1 (2010), 11-48.

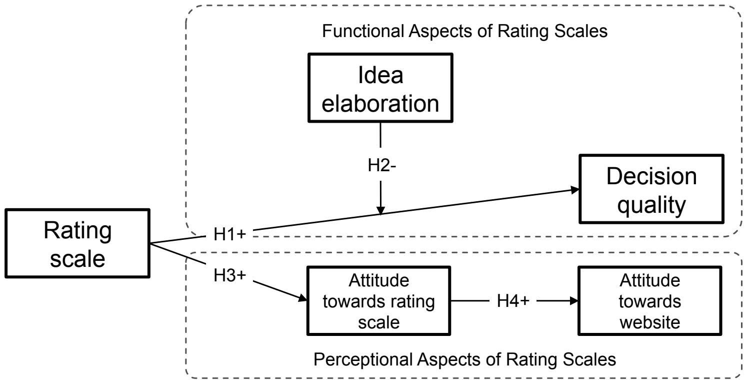

Figure 1 Research Model

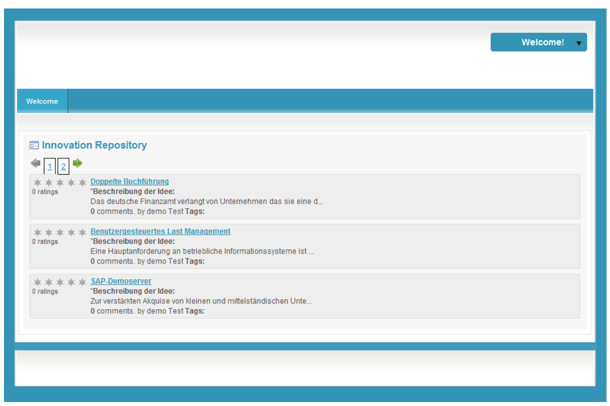

Figure 2 Single-criteria rating scale: ideas are evaluated in one dimension from one to five stars

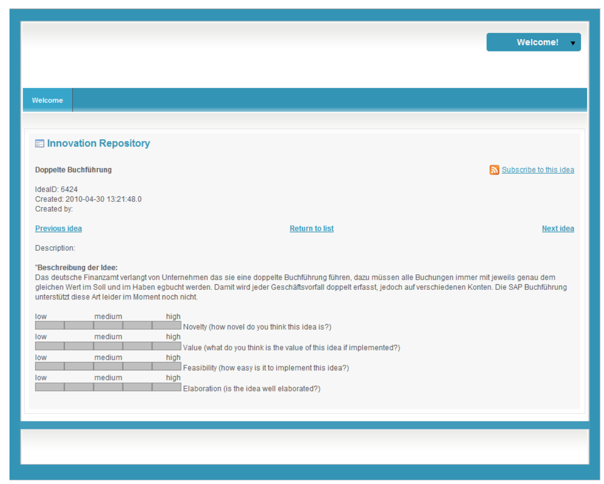

Figure 3 Multi-criteria rating scale: four 5-point scales for (1) novelty, (2) value, (3) feasibility, and (4) elaboration ranging from “low” to “high”.

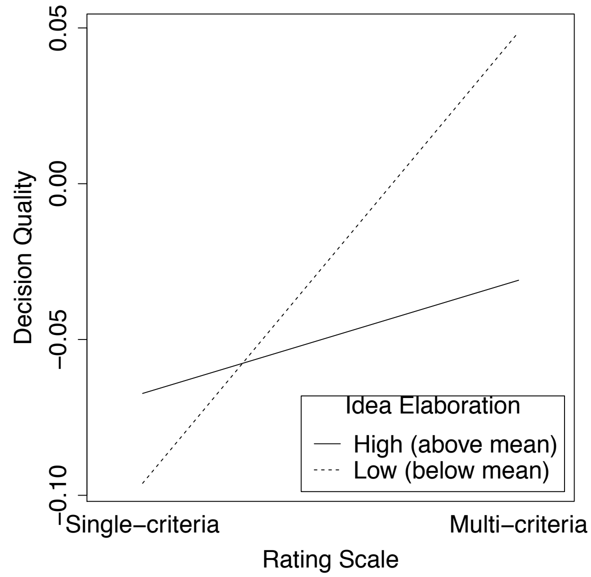

Figure 4 Interaction Effect of Rating Scale and Idea Elaboration

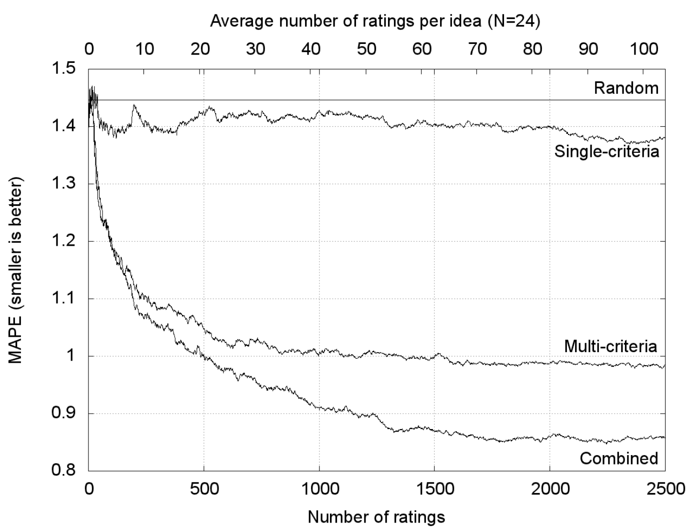

Figure 5 Simulation Results: Plot of the average MAPE of 100 simulation runs for the single-criteria and multi-criteria rating scales as well as a combination based on the idea elaboration moderator. For reference, the MAPE of a random shuffling is indicated at the 1.45 mark. The x-axis shows marks for both the total number of ratings drawn and the average number of ratings per idea (i.e., total ratings drawn divided by 24). The combined results use the single-criteria scale for highly elaborated ideas and the multi-criteria scale for lowly elaborated ideas.

Table 1. Items Measuring Attitude towards the Rating Scale and towards the Website

Attitude towards the Rating Scale		**Attitude towards the Website**	
Using the rating scale was…		Using the website was…	
ATR1	Dull – Exciting	**ATW1**	Dull – Exciting
ATR2	Not entertaining – Entertaining	**ATW2**	Not entertaining – Entertaining
ATR3	Negative – Positive	**ATW3**	Negative – Positive
ATR4	Frustrating – Satisfying	**ATW4**	Frustrating – Satisfying

Table 2. Factor Analysis of Questionnaire Items Measuring User Attitude

Item	**Factor**		**Cron-bach's α**	**Individual Item Reliability**	**Composite Reliability**	**AVE**
	Attitude Towards Rating Mechansims (1)	**Attitude Towards Website (2)**				
ATR4	0.84	0.11	0.79	0.52	0.79	0.49
ATR1	0.75	0.23		0.51		
ATR3	0.74	0.30		0.56		
ATR2	0.69	0.26		0.41		
ATW1	0.23	0.83	0.76	0.63	0.77	0.52
ATW2	0.19	0.80		0.46		
ATW3	0.27	0.74		0.48		
Eigenvalues	3.5	1.05				
Variance Explained	49.95	14.93				

MSA = 0.81; Bartlett-test of specificity: χ^2 = 528,327, p = 0.000; principal component analysis; varimax-rotation; n = 219. The bold values indicate the attribution of the variables to one of the three factors.

Table 3. Comparison of Expert Rating and Rating Scales

	Experts	Single-criteria	MAPE Unweighted mean aggregation	% improvement over single-criteria[a]	MAPE Response data-based weighted mean
Single-criteria	0.02	-	1.43	-	1.50
Multi-criteria	0.47**	0.22	1.02	40%**	0.95

N = 24, *** significant with $p < 0.001$, ** significant with $p < 0.01$

a Percentage of improvement over single-criteria rating scale = [MAPE single-criteria – MAPE instrument]/MAPE instrument (two-tailed paired t-test for difference).

Table 4 Testing the Moderating Effect of Idea Elaboration (H3)

Step	Independent Variable	B	R^2	ΔR^2
1	Idea Elaboration (Characters)	0.42*	0.00*	-
2	Idea Elaboration (Characters)	0.46	0.0**	0.00
	Rating Scale (Dummy)	0.19**		
3	Idea Elaboration (Characters)	1.17**	0.01**	0.01**
	Rating Scale (Dummy)	1.61**		
	Idea Elaboration x Rating Scale (Dummy)	-1.21*		

N = 3472, *** significant with $p < 0.001$, ** significant with $p < 0.01$, * significant with $p < 0.05$

Table 5. Testing the Mediating Effect of Attitude toward the Rating Scale (H3 and H4)

Step	Independent variable	B	β	R^2
1	Outcome: Attitude toward the Rating Scale			
	Predictor: Rating Scale (Dummy)	0.26*	0.13*	0.02*
2	Outcome: Attitude towards the Website			
	Predictor: Rating Scale (Dummy)	0.11	0.05	0.00
3	Outcome: Attitude toward the Website			0.29***
	Mediator: Attitude toward the Rating Scale	0.54***	0.54***	
	Predictor: Rating Scale (Dummy)	-0.03	-0.02	

N = 219, *** significant with $p < 0.001$, ** significant with $p < 0.01$, * significant with $p < 0.05$