Please quote as: Riedl, C.; Blohm, I.; Leimeister, J. M. & Krcmar, H. (2010): Rating Scales for Collective Intelligence in Innovation Communities: Why Quick and Easy Decision Making Does Not Get it Right. In: 30. First International Conference on Information Systems (ICIS) 2010, St. Louis, MO, USA.

RATING SCALES FOR COLLECTIVE INTELLIGENCE IN INNOVATION COMMUNITIES: WHY QUICK AND EASY DECISION MAKING DOES NOT GET IT RIGHT

Completed Research Paper

Christoph Riedl
Technische Universität München
Boltzmannstr. 3, 85748 Garching,
Germany
riedlc@in.tum.de

Ivo Blohm
Technische Universität München
Boltzmannstr. 3, 85748 Garching,
Germany
ivo.blohm@in.tum.de

Jan Marco Leimeister
Universität Kassel,
Nora-Platiel-Str. 4, 34127 Kassel,
Germany
leimeister@uni-kassel.de

Helmut Krcmar
Technische Universität München
Boltzmannstr. 3, 85748 Garching,
Germany
krcmar@in.tum.de

Abstract

The increasing popularity of open innovation approaches has lead to the rise of various innovation platforms on the Internet which might contain 10.000s user-generated ideas. However, a company's absorptive capacity is limited regarding such an amount of ideas so that there is a strong need for mechanism to identify the best ideas. Extending previous decision management research we focus on analyzing effective idea rating and selection mechanisms in online innovation communities and underlying explanations. Using a multi-method approach our research comprises a web-based rating experiment with 313 participants evaluating 24 ideas from a real-world innovation community, data from a survey measuring rating satisfaction of participants, and idea ratings from an independent expert jury. Our findings show that, despite its popular use in online innovation communities, simple rating mechanisms such as thumbs up/down rating or 5-star rating do not produce valid idea rankings and are significantly outperformed by the multi-attribute scale.

Keywords: Open innovation, absorptive capacity, rating, decision making, idea evaluation, collecting intelligence

Introduction

In the twentieth century, many leading companies generated and commercialized ideas for innovations mainly through in-house R&D laboratories. Today, companies are increasingly rethinking the fundamental ways of managing their innovation activities and overcoming their companies' boundaries in order to open up to other sources of innovation, which has become increasingly important. In this context, customers are seen as one of the biggest resources for innovations (Chesbrough 2006; Chesbrough et al. 2006; Enkel et al. 2005; von Hippel 1988; von Hippel 2005). Companies, no matter if they sell products or services, increasingly open not only their innovation process but also their production and sales process to customers and suppliers. Open innovation and crowdsourcing are thus gaining track in research and practice (Leimeister 2010; Leimeister et al. 2009). Positive impact of customer integration on company success and other measures have been demonstrated in various open innovation related research (e.g., Enkel et al. 2005; Gassmann 2006; Lakhani et al. 2007; Ogawa et al. 2006; von Hippel 2005; West et al. 2008).

The increasing popularity of open innovation approaches has led to the rise of various innovation platforms on the Internet (Riedl et al. 2009). Prominent examples are Dell IdeaStorm or MyStarbucksIdea, both comprising far more than 10,000 user-generated ideas. For putting these ideas into action, the most promising ideas have to be identified, must survive internal feasibility and profitability analyses, and be implemented in subsequent development projects. Thus, companies must develop appropriate organizational structures, processes, and routines for open innovation (Dahlander et al. 2010). However, even if capable structures exist, the host organization's absorptive capacity is limited as its employees cannot cope with the high dynamics of open innovation communities due to constraints of time and cognitive resources (Cohen et al. 1990; Di Gangi et al. 2009). In this context, collective decision making of many individual evaluations of community members could facilitate the process of identifying the best ideas. As the evaluation of external information is a central facet of absorptive capacity (Torodova et al. 2007) the application of these mechanisms may be a fruitful approach for enhancing the ability of incorporating customer generated innovation ideas.

Open innovation platforms generally use different rating scales that allow users to rate the submitted ideas. The effective design of those rating mechanisms enhances the validity and reliability of resulting idea ratings and supports the selection of the best ideas for further refinement or implementation. To date there is a lack of systematic study of how online communities can be exploited to better achieve different objectives of companies' innovation initiatives. Without such knowledge, ad-hoc use of online communities may result in inefficient resource utilization and may impair the effective integration of customers into the innovation process. This problem may be particularly salient with the increasing choices and sophistication of tools available for the creation of online communities. In addition to the effective design of those rating mechanisms users' satisfaction with the website constitutes an important antecedent of successful community building. Much of the research in the HCI literature frequently excludes affective variables such as attitudes (satisfaction) from system evaluation. However, attitude measures have been used as surrogates for success at different levels of granularity (Galletta et al. 2004). The amount of satisfaction a user has with the Web site's interface (an attitude) is seen as a dominant component of a general attitude about returning to the site and thus for successful community building (Cyr et al. 2009; Galletta et al. 2004).

Extending previous decision management research, we focus on analyzing effective idea rating and selection mechanisms in online innovation communities and underlying explanations. This research seeks to advance knowledge about the effective and efficient utilization of information technology for the improvement of online innovation portals. To gain insights into how different rating mechanisms work we conducted a multi-method study. Using a pool of 24 real-world ideas submitted in a public idea competition (Blohm et al. 2010) our study comprised a web-based experiment, a survey measuring rating satisfaction of participants, and an independent expert (n=7) rating of idea quality. Through triangulation, we seek to gain a more comprehensive insight into how community rating mechanisms work. We use an experimental design for comparing rating scales to judge idea quality with different granularity. These scales comprise a binary rating (thumbs-up, thumbs-down), a 5-star rating, and a complex rating involving four 5-star scales reflecting the different traits of idea quality grounded in creativity and innovation management research.

In summary, the research has the following goals:

1. From a theoretical perspective, we create and test a model to analyze the influence of the rating scale on rating quality and user satisfaction. Thus, our paper provides a first experiment validating different rating scales for community evaluations.
2. From a methodological perspective, the research uses three different methods to analyze and interpret the validity and effectiveness of three different rating scales commonly used in online innovation portals. The outcomes of three methodologies using web-based rating experiments, a questionnaire, and independent expert ratings are compared and investigated to understand the research model of this study.
3. From a practical perspective, our research provides actionable design guidelines for community-based rating mechanisms in innovation portals. Following these design recommendations, community ratings in innovation portals should be improved.

The paper is structured as follows. We first present our research model and develop relevant hypothesis. We then present our research methodology including detailed description of the experimental task and design. The remainder of the paper then presents the results of the experiment, the questionnaire, and the expert rating followed by a discussion of the results. Finally, the conclusion discusses limitations and opportunities for future research.

Theoretical Background

Idea Quality

Since all innovation begins with creative ideas (Kristensson et al. 2004), the evaluation of new product ideas is strongly related to the assessment of their inherent creativity. Creativity and idea quality are both complex constructs that have been a subject for creativity, group support system and innovation researchers for years. In the context of customer-generated new product ideas, idea quality consists of four distinct dimensions: novelty, feasibility, strategic relevance and elaboration (Blohm et al., 2010).

Creative solutions are generally characterized as being new and useful (Amabile 1996; Mayer 1999; Niu et al. 2001; Plucker et al. 2004). Novelty is often defined as something being unique or rare. In this context, new ideas have not been expressed before (MacCrimmon et al. 1994). A closely related trait of novelty is originality. Original ideas are not only new, but also surprising, imaginative, uncommon or unexpected (Ang et al. 2000; Dean et al. 2006), and many researchers see originality as the most important facet of creativity (Besemer et al. 1999; Runco et al. 1999; Walcher 2007). Another attribute of novelty is the paradigm relatedness (Besemer et al. 1986; Finke et al. 1996; Nagasundaram et al. 1994). This refers to an idea's transformational character, and describes the degree to which an idea helps to overcome established structures, i.e., how radical or revolutionary it is (Besemer et al. 1986; Christiaans 2002). From a new product development perspective, an idea's paradigm relatedness refers to its innovativeness.

However, an idea's novelty is not sufficient for being unique and useful. Usefulness is the extent to which the idea responds to or solves a problem that is tangible and vital (Amabile 1996; Dean et al. 2006). This dimension is also called an idea's value or relevance (Dean et al. 2006; Kristensson et al. 2004; MacCrimmon et al. 1994). In the scope of new product development, this refers frequently to an idea's financial potential (Cady et al. 1999; Franke et al. 2006; Lilien et al. 2002; Rochford 1991), the strategic importance in terms of enabling competitive advantages (Cady et al. 1999; Lilien et al. 2002; Rochford 1991), as well as the customer benefit that an idea endows (Ogawa et al. 2006; Walcher 2007). From the innovator's perspective, an idea's feasibility is another vital dimension of idea quality. This dimension captures the ease with which an idea can be transformed into a commercial product and the fit between the idea and the organizer (Cady et al. 1999; Lilien et al. 2002; Rochford 1991). In this context, the fit is two-pronged: From an internal perspective, it refers to the organizer's strategy, capabilities and resources, and from an external perspective, to the fit between the idea and the organizer's image. Another trait of a high quality idea is its elaboration, which can be seen as the extent that it is complete, detailed and clearly understandable (Dean et al. 2006). Furthermore, this refers not only to an idea's description but also to its maturity (Franke et al. 2006).

Customer-Generated New Product Ideas

Generally, new product ideas are creative products which combine existing elements in a novel manner and satisfy pre-existing criteria such as a firm's strategy, its customers and its competitors. The ideas are the result of a non-deterministic creative process and yield semantic information that overlaps the information in the initial knowledge (Johnson-Laird 1993). Customer-generated new product ideas may be of great value for a company as they provide novel information about customer needs (need information) and new ways of fulfilling these needs (solution information) that have hitherto not been considered by the company (von Hippel 1994). However, these ideas are often not very specific and show a rather low degree of elaboration and maturity. Usually, they have not been revised (Blohm et al. 2010). Thus, customer-generated new product ideas are often vague and blurry. Moreover, the pre-existing structures the ideas have to cope with have usually not been taken into account in the idea generation process of the customers.

The Decision Process

Rating ideas in open innovation platforms, community members run through a cognitive process that is very comparable to the one of responding to a survey. Like in answering survey questions, community members have to understand the idea in the first instance, to make an individual judgment about an idea's quality as well as to perform a rating on a given rating scale in order to express their judgment of idea quality. Thus, it can be assumed that community members undergo a decision process that involves four basic steps (Tourangeau et al. 2000):

1. **Comprehension** encompasses attending to the idea and accompanying instructions, assigning a meaning to the surface form, and inferring the idea's point (Tourangeau et al. 2000). In this process, idea evaluators initially see the length of the idea and estimate the effort to evaluate the idea's quality. Then, they assess the form of the idea and the meaning of illustrations and other visual design elements. In the third step the idea is read and the meaning of the words assessed (Ganassali 2008).
2. **Information retrieval** involves recalling relevant information from long-term memory and bringing it into an active state, in which it can be used to rate the quality of the ideas. This process includes the adoption of a retrieval strategy, using cues to trigger the recall of information, remembering generic and specific memories and filling in missing details through inference (Collins 2003; Tourangeau et al. 2000).
3. **Judgment** is the process in which respondents formulate their answer to the idea rating task (Collins 2003). In this process the retrieved information are evaluated regarding completeness and relevance and integrated into an overall judgment (Biemer et al. 2003; Tourangeau et al. 2000). According to Tourangeau et al. (2000) information integration is an iterative process in which the retrieved information is evaluated regarding the idea at hand. This initial judgment is then altered in respect to the evaluations of the following information. The final judgment can then be seen as an average of the evaluations of the retrieved information.
4. **Reporting and response selection:** In this last step, respondents map their judgment onto the given response options. The respondents convert their judgments into close-ended items with an ordered set of response categories which are mapped to the traits of the idea they have to rate. The most extreme facets of idea quality or its sub dimensions are mapped to the scale endpoints serving as anchors for the remaining scale points; ideas of intermediate quality are then mapped in the middle of these two bipolar extremes (Tourangeau et al. 2000). After this process of encoding, the final records are formed. However, after this mental record formation, the response may be altered in respect to consistency with previous responses, social acceptability or other influencing factors (Biemer et al. 2003).

However, this decision process does not have to be a linear one. In rating ideas the decision makers can spring back to subsequent stages and run through the following ones in an iterative fashion (Biemer et al. 2003; Tourangeau et al. 2000).

Hypotheses and Model Development

Generally, a rating scale's complexity and its optimal number of categories are depending on the ability to differentiate a specific circumstance as well as the respondent's ability to discriminate the given circumstance

(Malhotra 2007). Prior research shows how the number of response alternatives affect the psychometric properties of a scale and most researchers found an increasing granularity of the scales to positively influence the reliability and the factorial validity of the complex constructs that have been measured (Ferrando 2000; King et al. 1983; Lozano et al. 2008).

On the one hand, this derives from statistical effects such as variance amplification that comes along with more granular rating scales (Malhotra 2007). Due to more response options the answers of the responders will be spread more widely leading to better psychometric properties of the scale. On the other hand, the presentation of the question or the idea in an idea rating task - including the rating scale - is one of the most important variables that may affect the behavior of respondents (Ganassali 2008; Tourangeau et al. 2000). Respondents act as cooperative communicators, and they will endeavor to make sense of the questions by drawing on all information including formal features, such as the numeric values of rating scales or the scales' graphical layout (Schwarz 1996). This is especially true, when respondents are unsure about what is being asked and have to answer tough questions with no 'right' answer like rating idea quality (Christian et al. 2004). Thus, an appropriate design of the rating scale can facilitate the process of mapping the response onto a given scale. Research suggests that the minimum number of categories for ensuring an appropriate level of reliability is four (Lozano et al. 2008). The dominant design (Utterback 1996) of current innovation portals using only a binary scale violates this recommendation and basically introduces a dichotomous format measured by only a single item (promote idea and demote idea). In this context, evaluating the quality of customer-generated new product ideas could be oversimplified with a binary scale. Given that the overwhelming majority of innovation portals use this binary rating mode we see a need for testing it regarding its suitability in measuring idea quality compared to the other scales.

A more complex scale, like a 5-star rating scale, may better support the process of integrating the different aspects of the idea into a single judgment and mapping this on different categories of the rating scale. Moreover, rating scales embodying cues such as definitions that explain the meaning of uncommon words may help the respondents to better express their ratings (Christian et al. 2007; Conrad et al. 2006) as the task can be better understood and subsequently more relevant information can be retrieved for the judgment. Thus, it is likely that a rating scale that breaks down the complex construct idea quality in different sub-scales addressing the different aspects of idea quality together will yield a higher rating accuracy than single item rating scales. Summing up these considerations we assume:

H1: The granularity of the rating scale positively influences its rating accuracy.

Contrary to an apparent weakness of the hypothesis, we see a need to test the influence of rating scale granularity on the accuracy of user ratings due to (1) the well accepted strong effect of rating scales on respondent behavior in general; (2) the dominant design of innovation platforms using only a single item, binary scale which constitutes a special case of a rating instrument which has not been systematically studied regarding its suitability in measuring idea quality; and (3) the aim of developing design recommendations regarding the optimal granularity of rating scales in the context of innovation platforms.

For community operators, a scale's rating accuracy is not the only criterion that has to be considered when a rating mechanism is designed. A continuous usage of the rating scales is depending on how the community members perceive the rating process (cf. Ebner et al. 2009). Generally, satisfaction has an evaluative focus, reflecting how favorable or unfavorable a person is toward a specific alternative (Fishbein 1966). In contrast to post-purchase satisfaction that requires experience with the consequences of the chosen product, post-decision satisfaction arises frequently immediately after the decision (Sainfort et al. 2000).

According to Janis and Mann's (1977; 1982) conflict theory of decision making, post-decision satisfaction is heavily influenced by decisional stress that comes along with emotionally-laden decisions. Highest decision satisfaction is perceived in decision situations with an intermediate degree of stress as this indicates a conflict that could successfully be solved by the decision maker. Evidence from the neuropsychological literature suggests that cognitive judgments are generally preceded by emotional ones (Goleman 1996; LeDoux 1998; Zajonc 1980). This form of experience that we call 'feeling' accompanies all cognitive cognitions. In the context of decision processes these emotions arise already in the comprehension and information retrieval processes before the actual judgment is done (Biemer et al. 2003). Thus, judgments of objective properties are often influenced by affective reactions (LeDoux 1998; Zajonc 1980).

Binary rating scales force respondents to make a distinct decision about an idea's quality. However, an idea's quality as well as the emotions that arise during the decision making process are not likely to be dichotomous so that

respondents may fail to map all facets of their judgment onto the two response options. In this situation, the discrepancy between their affective and cognitive evaluation may lead to a conflict situation with high decisional stress. The idea raters may perceive a state of 'hyper-vigilance' (Janis et al. 1977; 1982), a mismatch between the expectations of rating the ideas accurately and the perceived quality of their rating. Thus, we assume that more granular rating scales such as a 5-star rating and the complex rating lead to a higher rating satisfaction as they elicit less stressful, intermediate conflict situations that idea raters can better cope with:

H2: The granularity of the rating scale positively influences the users' satisfaction with their ratings.

However, in order to derive sound design guidelines for rating scales harnessing the wisdom of community members, not only rating accuracy and rating satisfaction have to be taken into account but also contextual factors. In particular, different rating scales might be applicable for different user types. For instance, it might be an intuitive design pattern to provide less knowledgeable users with a different scale than more knowledgeable expert users. This implies that a moderating effect between the rating scales and the level of user expertise has to be assumed. Generally, a moderating effect occurs when the relation between two variables is dependent on a third one, which alters the direction or the strength of the relationships between the other ones (Baron et al. 1986; Frazier et al. 2004). Focusing on the moderating effect rather than a direct effect allows us deriving detailed design recommendations regarding the IT artifact which is a key aim within IS research (Benbasat et al. 2003).

Among creativity researchers there is a broad consensus about the fact that experts with a high degree of expertise in the given domain are appropriated best for evaluating the quality of creative products (Amabile 1996; Caroff et al. 2008). This holds true for innovation management where new product ideas are generally evaluated by a small team of interdisciplinary experts (Toubia et al. 2007). Referring to the idea rating process, a high degree of expertise should enable respondents to better comprehend the evaluation task as the assimilation of new information is facilitated when already existing mental structures can be used to process the information (Sudman et al. 1996). Following this argumentation, it will be easier for more knowledgeable raters to integrate the different aspects of their decision on a given rating scale than for less knowledgeable users. In other words, expert users might be better able to adequately express their quality judgments on a binary, thumbs up/down scale than less experienced ones. As the decision process is rather iterative than linear, the rating scale will influence the weighing of the different traits of idea quality and the potential tradeoffs between them. This process will be easier for expert users yielding more accurate results. Moreover, a more complex multi-attribute scale may engage raters to reflect about the specific traits of idea quality that have to be judged in order to make a sound quality assessment. This effect may be greater for less knowledgeable users. For them the single criteria of the complex rating scale may provide bigger hints as it is likely that they would not have thought about them with a less granular, single dimension scale. In comparison, expert users may benchmark the idea against a broad range of alternative solutions, so that even with less granular scales different traits of idea quality will be considered implicitly. Summarizing these theoretical considerations, we assume that user expertise should have a moderating effect on the relationship between the used rating scale and the accuracy of customer ratings:

H3a: User expertise moderates the relationship between rating scale granularity and rating accuracy such that the positive relationship will be weakened for high levels of user expertise and strengthened for low levels of user expertise.

Confidence is a key variable in decision making and in judgmental tasks. It can be described as the belief in the accuracy of one owns decision (Sniezek 1992). Though confidence and decision satisfaction are strongly associated, they are conceptually distinct as confidence is a belief and satisfaction an attitude (Sniezek 1992) which can generally be defined as a function of salient beliefs (Fishbein 1966). Thus, confidence is often conceptualized as a predictor for decision satisfaction (Keller 1983; Small et al. 2000). Generally, high confidence is grounded in thorough cognitive information processing that underpins the individual decision process. Moreover, experts were generally found to have a higher confidence in their decisions as they are aware of their expert status and are convinced of their expertise (Tetlock 2006). Consequently, raters having a low expertise are likely to be less confident and thus being less satisfied with their results. However, knowledgeable and less knowledgeable raters may perceive the rating scales' granularity in different ways. The decisional stress that comes along with the forced decision of the binary rating scale may be higher for knowledgeable raters as more relevant information is activated that have to aggregated in a single response undergoing the decision process. Thus, the state of hyper-vigilance might be more pronounced for high expertise raters than for low expertise raters, who have to integrate a smaller amount of information only. Additionally, less knowledgeable users facing a more complex, multi-criteria rating scale may experience a feeling of greater insecurity than with a less granular scale. The single rating criteria ask the

user for specific information he or she may not able to provide and thus creating decisional stress impairing decision satisfaction. Summing up, we assume the relationship between the rating scale and rating satisfaction to be moderated by the level of user expertise of the rater:

H3b: User expertise moderates the relationship between rating scale granularity and rating satisfaction such that the positive relationship will be strengthened for high levels of user expertise and weakened for low levels of user expertise.

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

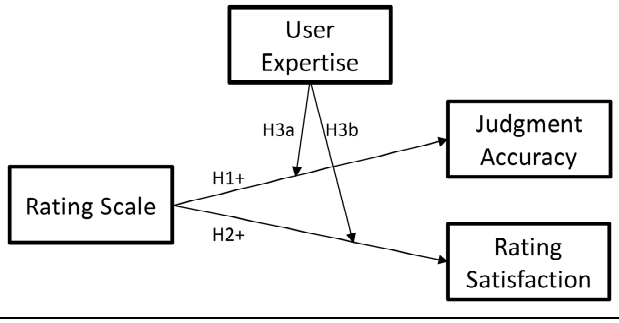

Figure 1. Research Model

Research Methodology

Participants

People participating in topic related open innovation platforms and virtual communities can be seen as the target population of our experiment and open innovation communities in general. Prior research has shown that people engaged in user innovation and virtual communities for innovation are predominantly male, young and well educated (Franke et al. 2003; Jeppesen et al. 2006; Jokisch 2007; Kristensson et al. 2004; Schulz et al. 2008; Walcher 2007). 349 participants took part in the experiment of those 313 were included into the analysis. Our sample population consisted of undergraduate and graduate students from four information systems courses, two of them directly related to SAP education, as well as research assistants from the same area at a large German university. Students from three of the courses were offered homework credit points for participating in the experiment. There was no significant differences between rewarded and none rewarded students.

We considered students of the selected SAP related educational courses and information system experts to be appropriate subjects for this study because the experimental task requires knowledge of SAP software systems to judge idea quality related to SAP software. Furthermore, it can be argued that IS/SAP course students are suitable experiment participants as they represent actual users of innovation platforms. On a general level, Voich (1995) found the values and beliefs of students to be representative of individuals in a variety of occupations. Table 1 summarizes the demographic profile of the study participants.

Table 1. Participant Demographics

	313
Mean age	22.81 years
Gender	Male: 67.7 % Female: 32.3 %
Highest University Degree	None (high school only): 69.3 % Bachelor: 25.2 % Master: 5.4 %

Idea Sample

The ideas evaluated in this experiment were taken from an idea competition that was conducted in summer 2008 with a runtime of 14 weeks (Blohm et al. 2010). In this idea competition SAP users were asked to submit ideas that improve the SAP software or that bring out radical innovations in the scope of the SAP software. In total 58 new product ideas were contributed by 39 different users.

Among these ideas, idea quality is normally distributed. The ideas varied in length between half and a full A4 page. Conducting an experiment with all ideas implied a substantial workload for all experimentees. Hence, a stratified sample of 24 ideas was drawn in order to maximize participation. This sample comprised 8 ideas with high, medium and low quality respectively. The sample size was considered sufficient as 20 to 30 ideas are generally used to measure the variance of creativity ratings in creativity research (Caroff et al. 2008; Runco et al. 1993; Runco et al. 1992).

Experimental Task and Design

The experiment has been performed as a web-based experiment using a standard innovation portal developed by the authors. Standard features of the platform like idea submissions, commenting, searching and sorting have been disabled and only the rating mechanisms were activated (see screenshots in Appendix A). The order of ideas on the platform has been randomized for each user so that all participants evaluated the ideas in a different order and a position bias can be avoided (Malhotra 2007). In this regard the userID served as the random seed. The ideas to be evaluated comprised of a title and a description. Participants performed the task on their own computers (at home, at work, in a computer lab) via a web browser. Before starting the experiment we tested whether all common browsers displayed the innovation portal in a similar way and no irregularities were discovered. As a web experiment closely reflects the actual usage scenarios of virtual communities for innovation and open innovation platforms, a high external validity of our results can be assured. Participants can rate the ideas in their natural environment and can allocate as much time to completing the rating task as they want to. Furthermore, the internal validity of results is assessed by analyzing the log files on the idea platform. Doing so, user responses that have an improbable response behavior such as responding too fast can be identified and excluded from analysis. The time stamp of each performed rating has been recorded so as to identify users who just clicked through the rating in order to exclude them from the sample. Every idea is rated individually by one of three scales (refer to Appendix A).

The system provides immediate visual feedback to a successful rating (i.e., the respective button/star is highlighted). Users are also able to update their ratings again with immediate visual feedback. Through the update mechanism it is assured that every user can rate every idea only once. In order to avoid information cascades (Easley et al. 2010) and thus a rating bias deriving from other participants' ratings, rating information of other participants is not visible. Ideas that have not been rated are clearly visible due to the colored highlighting that is shown once an idea has been rated. This made it convenient for users to navigate through the system to identify ideas that have not yet been rated or to check for completeness.

Participants were asked to rate the ideas with the following task description:

Please carefully read through all ideas and provide a rating of the idea quality as judged by your personal experience. Please consider an idea's overall quality in terms of its novelty, relevance, feasibility and elaborateness for your rating as indicated by the idea's title and description.

Rating Scales

For our experiment three different configurations of the innovation platform have been set up, one for each of the rating scales. Each system was accessible under a different URL. The scales comprise of a binary rating scale ("promote/demote rating"), a five-point rating scale ("5-star rating") and a complex rating scale. Whereas the promote/demote as well as the 5-star rating reflect an aggregated measure for idea quality, the complex rating scales consisted of four 5-point rating scales reflecting the single dimensions of idea quality used in the expert rating (Table 2). The 5-point rating scale of the complex rating ranged from "low", through "medium" to "high" (cf. Appendix A). In order to avoid confounding effects of respondent fatigue and satisficing (Tourangeau et al. 2000), we reduced the single items of the expert evaluations to the four main dimensions.

Table 2. Rating Aspects of the Complex Rating	
Rating attribute	Label with rating instruction
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?
Elaboration	Is the idea well elaborated?

Procedure

Participants of the sample population were first randomly assigned to one of the three ratings scale treatments (random sampling without replacement). Based on the random assignment, we invited the participants via a personalized email including a link with the respective system URL and the online questionnaire. Participants completed the rating task distributed over the experiment duration of four weeks (November and December 2009). After the four weeks the online systems were closed and the data sample was exported for the data analysis. Table 3 summarizes the participants for each of the three rating scale treatments.

Table 3. Experimental Design			
	Promote/Demote	5-Star	Complex Rating
N	94	103	116

A Multiple Method Approach

In this study, three research and analysis methodologies are employed (web experiment, quantitative survey analysis, expert rating) to investigate our hypotheses. Various researchers advocate the use of multiple methods of data collection, both to gain a deeper insight and more reliable results (Boudreau et al. 2001; Palvia et al. 2004; Sharma et al. 2009). Similar to an approach taken by Cyr (2009) we aim for greater robustness in the current investigation through the use of multiple methods.

Experiment Rating

Initially, 349 participants took part in the experiment. Idea raters that did not rate all ideas, did not fill out the survey completely or rated the ideas in less than 5 minutes were discarded form the analysis. The remaining 313 idea raters performed 15864 ratings in total. The median time it took the users to rate the 24 ideas (measured by the difference between the timestamp of the first and the last rating that a given user submitted) was 35 minutes and 35 seconds. It has to be noted, however, that the time taken for submitting the ratings does not include the time a user spent on reading through the ideas (i.e., a user might spend a considerable amount of time reading through all ideas before starting to submit ratings).

Questionnaire

User expertise and rating satisfaction were collected conducting an online survey among the participants after the experiment. The scales for measuring expertise and satisfaction were adapted from scales that have already been used in the context of open innovation and computer-human interaction studies before. All items were measured with a 5-point Likert scale.

According to Lüthje (2004) user expertise that is relevant for user innovation involves two distinct facets: product-related knowledge and use experience. Product-related knowledge consists of know-how about the architecture of the product, the used materials and the underlying technology. Use experience sprouts from frequently using a given product. We developed our user expertise scale based on product-knowledge (Bloch et al. 1989; Flynn et al. 1999;

Lüthje 2004) and use experience scales (Griffin et al. 1996; Spann et al. 2009) that have already been used in an open innovation context.

Satisfaction with rating scales is usually not measured as a quality criterion of rating scales (Ganassali 2008). Thus, we developed our scale for measuring rating satisfaction based on scales for measuring satisfaction with websites (Oliver et al. 1989; Shankar et al. 2003) and website usability as it strongly determines satisfaction in human computer-interaction (Lindgaard et al. 2003; Shankar et al. 2003).

The entire survey was pretested with a small sample of ten participants, reflecting the different groups of experimentees. They were asked to provide detailed comments on the survey such as working or concept confusion. Based on this feedback minor changes to the survey were made.

Expert Rating

For assessing the validity of the different rating scales the participants idea quality ratings derived with these scales are compared with an independent expert rating. The ideas from the idea contest were evaluated by a qualified expert jury using the consensual assessment technique (Amabile 1996). This assessment technique derives from creativity research and was already used several times for assessing the quality of customer generated new product ideas (Blohm et al. 2010; Franke et al. 2006; Kristensson et al. 2004; Matthing et al. 2006; Piller et al. 2006; Walcher 2007). Using this method ideas are evaluated by a jury consisting of experts in the given domain. In our case the jury consisted of 7 referees, which were either university professors, employees of the initiator SAP or the German SAP University Competence Centers. The complex construct of idea quality was operationalized in four dimensions and measured in 15 items. For evaluation the idea descriptions were copied into separate evaluation forms which contained the scales for idea evaluation as well. The evaluation forms were handed out to the referees in a randomized order. All judges were assigned to rate the ideas with the 15 different items on a rating scale from 1 (lowest) to 7 (highest). Each member of the jury evaluated the ideas independent from the others. In order to assess idea quality validly and reliably we conducted exploratory and confirmatory factor analysis. A detailed description of this procedure can be found in Blohm et al. (2010).

Analysis

Construct Validation

In the first instance, we tested the reliability and the validity of the satisfaction and the user expertise scales. The means, standard deviations, and the intercorrelation of these variables are depicted in Table 4. Performing exploratory factor analysis with SPSS 17.0 we tested their dimensional structure. All items loaded unambiguously on the 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 (Malhotra 2007). 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 (Malhotra 2007). With Alphas of at least 0.68 this criterion can be considered as met.

Table 4. Means, Standard Deviations, and Intecorrelations of the Study Variables 1 and 2

Variable	Mean	Standard Deviation	1	2
User Expertise	1.71	0.43	-	
Satisfaction	3.36	0.85	-0.05	-

Subsequently, we tested these factors applying confirmatory factor analysis using Amos 17.0. 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 5). 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 (Bagozzi et al. 1988). The discriminant validity of the

Table 5. Factor Analysis of Idea Quality

Item	Factor		Cronbach's α	Individual Item Reliability	Composite Reliability	AVE
	User Expertise (1)	Rating Satisfaction (2)				
EXP1	**0.91**	-0.02	0.91	0.77	0.91	0.72
EXP4	**0.89**	-0.07		0.64		
EXP2	**0.87**	0.03		0.63		
EXP3	**0.87**	0.03		0.77		
SAT2	-0.04	**0.80**	0.68	0.33	0.75	0.51
SAT1	-0.05	**0.78**		0.39		
SAT4	0.07	**0.76**		0.45		
Eigenvalues	3.16	1.82				
Variance Explained	45.19%	26.04%				

MSA = 0.81; Bartlett-test of specificity: $\chi^2 = 1094.27$, p = 0.000; principal component analysis; varimax-rotation; n = 313. The bold values indicate the attribution of the variables to one of the two factors.

factors was checked by using the Fornell-Larcker criteria which claims that one factor's AVE should be higher than its squared correlation with every other factor (Fornell et al. 1981). The squared multiple correlation between the two factors is 0.02. As this is smaller than the AVE of both factors, discriminant validity can be assumed. For both factors, Individual Item Reliabilities were calculated. Two items of our satisfaction scale (SAT2, SAT3) violated the minimum threshold of 0.4 (Bagozzi et al. 1988). However, with a value of 0.33 the Individual Item Reliability of item SAT2 was only slightly below the threshold of 0.4. As this item was to be considered an important trait of satisfaction, it was not excluded from analysis. Only item SAT3 was excluded as it had a small Individual Item Reliability of 0.13. Overall, the scale's good reliabilities based on Cronbach Alpha can be confirmed.

Finally, we checked the global fit of our measurement model by conducting a Chi-Square (χ2)-test. The χ^2-test was not significant (p = 0.14) and the χ^2 / df-ratio was 1.43, well below the upper threshold of 5.00, which indicates good fit (Wheaton et al. 1977). Furthermore, global fit measures suggested excellent fit as well: GFI = 0.98 (Goodness of Fit Index; $\geq$ 0.9), AGFI = 0.97 (Adjusted Goodness of Fit Index; $\geq$ 0.9), NFI = 0.98 (Normed Fit Index; $\geq$ 0.95), CFI = 0.99 (Comparative Fit Index; $\geq$ 0.95), RMSEA = 0.04 (Root Mean Square Error of Approximation; $\leq$ 0.06) and SRMR = 0.03 (Standardized Root Mean Square Residual; $\leq$ 0.11) (Browne et al. 1993; Bühner 2008). Thus, the instrument was successfully validated using both exploratory and confirmatory factor analysis.

Hypothesis Testing

Generally, it can be assumed that participants' evaluations are of high accuracy if the participants are able to effectively identify the 'best' ideas among all ideas. In the context of open innovation communities, the best ideas would be those creating the highest profits after having been implemented by the company. However, this true idea quality is a priori unknown and the community ratings can only serve as a pre-selection for a further internal review phase (Di Gangi et al. 2009). Thus, the particular quality score of a given idea that has been assigned by the community is in principle not relevant. More important is that the best ideas are identified correctly by the participants (Reinig et al. 2007). In creativity research judgmental accuracy of laypersons is often 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 (Runco et al. 1993; Runco et al. 1992).

Current research about customer-generated new product ideas shows that about 10-30% of these ideas can be regarded as high quality ideas (Blohm et al. 2010; Franke et al. 2006; Walcher 2007). Thus, we defined two cut-off criteria with 5 ideas (ca. 21%) and 8 ideas (ca. 33%) from the high quality sample strata as "top ideas" as this corresponds to the ratio of high quality ideas in real-world settings. Respectively, the 5 and 8 ideas from the low quality sample strata were classified as "bad ideas." We performed all following analyses with both cut-off criteria

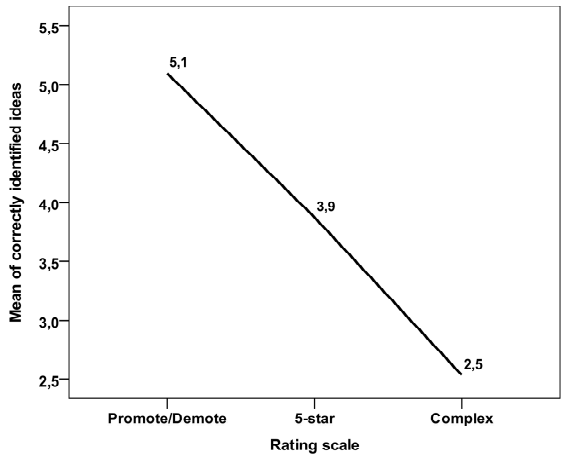

Figure 2. Correct Identification of Good and Bad Ideas

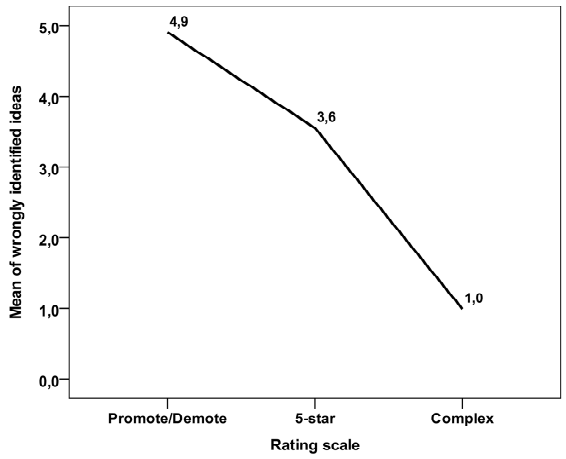

Figure 3. Error Identifying Top Ideas as Good and Bottom Ideas as Bad

leading to almost identical results. Hence, we report only the results of the more severe 5 idea cut-off-ratio as we think that this better reflects reality as it is likely that idea quality is concentrated among few good ideas. The individual user ratings of all rating scales were aggregated by calculating the arithmetic mean.

In the first instance, we tested the accuracy of our rating scales by counting the correctly classified high and low quality ideas of each user (cf. Figure 2 and 3). Analysis of Variance (ANOVA) revealed that the binary promote/demote rating yielded the significantly highest amount of correctly classified ideas ($F_{2,310} = 69.78$, $p < 0.001$). Bonferroni-post-hoc comparisons revealed that differences between all rating types are significant ($p < 0.001$). However, simultaneously the promote/demote rating leads to significant higher misclassification of ideas compared to the 5-star and the complex rating, so that good ideas are wrongly classified as bad ones and vice versa ($F_{2,310} = 225.14$, $p < 0.001$). The rating error is significantly different between all rating types ($p < 0.001$).

Thus, we operationalized rating accuracy with an adjusted Fit-Score, which was calculated by subtracting the wrongly classified ideas from the correctly classified ideas. The hypotheses H1 and H2 were tested applying Analysis of Variance (ANOVA). Separate analyses were run for rating accuracy and satisfaction as a dependent variable.

Significant main effects for the influence of the rating scale on rating accuracy ($F_{2,310} = 9.05$, $p < 0.001$) as well as satisfaction ($F_{2,310} = 4.52$, $p = 0.01$) could be found (cf. Table 6, Panel A). Thus, H1 and H2 can be supported. Post-hoc comparisons reveal that the complex rating scale leads to a significantly higher rating accuracy than the promote/demote rating and the 5-star rating ($p < 0.001$). Between the promote/demote and the 5-star rating scales no significant differences can be observed. Regarding H2, the 5-star rating leads to the highest degree of user satisfaction that is significantly higher than the satisfaction of promote/demote raters ($p = 0.01$). No significant differences between complex and 5-star rating could be found (cf. Table 6, Panel B).

We followed the recommendations of Frazier et al. (2004) and Cohen et al. (2003) and applied moderated, hierarchical OLS regression in order to test for the moderating effects of user expertise (Hypotheses H3a and H3b). As the rating scale has categorical measurement level with three levels, we had to recode it into two dummy variables. We applied the dummy coding scheme as suggested by West et al. (1996). In this coding scheme the binary rating scale served as reference group. Thus, the first dummy compares the 5-star rating and the second dummy the complex rating scale with the binary rating scale. As we used the factor scores for User Expertise there was no need of standardization. Subsequently, we estimated the following regression equation:

$$Y = b_0 + b_1 \textit{User Expertise} + b_2 \textit{ Dummy 1} + b_3 \textit{ Dummy 2} + b_4 \textit{ User Expertise x Dummy 1} + b_5 \textit{ User Expertise x Dummy2}$$

Table 6. ANOVA Results

Panel A. Effect of Rating Scale and Expertise on Rating Accuracy

Source	df	Sum of Squares	Mean of Squares	F	Hypotheses	Supported
Between Groups	2	121.23	60.61	9.05***	H1	Yes
Within Groups	310	2075.77	6.70			
Total	312	2196.99				

Panel B. Effect of Rating Scale and Expertise on Rating Satisfaction

Source	df	Sum of Squares	Mean of Squares	F	Hypotheses	Supported
Between Groups	2	7.44	3.72	9.05***	H2	Yes
Within Groups	310	253.36	0.82			
Total	312	270.80				

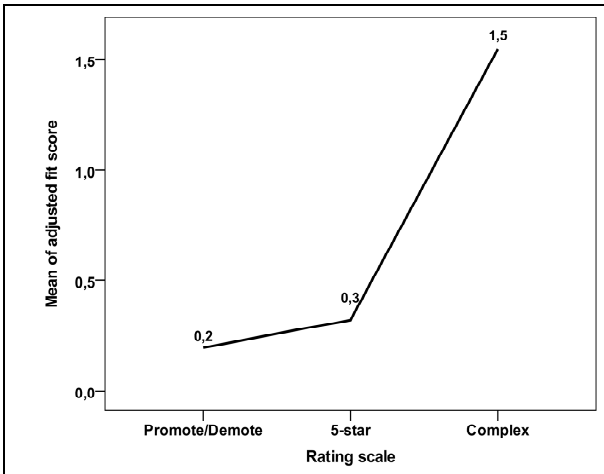

Figure 4. Rating Accuracy

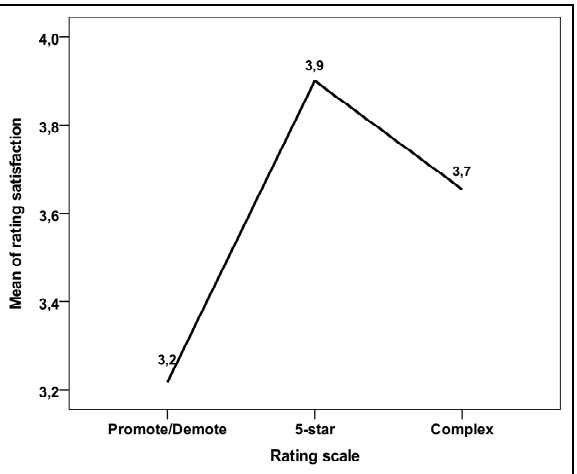

Figure 5. Rating Satisfaction

The moderator effect can be tested with a multiple degree of freedom omnibus F test representing the stepwise change of explained variance for the step in which the interaction terms are entered (Frazier et al. 2004; West et al. 1996). Including the two dummy variables into the regression equation a significant gain in explained variance can be detected for rating accuracy and rating satisfaction (cf. Table 7, Panel A). This is consistent with this paper's previous findings that rating scale granularity affects rating accuracy. However, no significant gain in explained variance can be found for the inclusion of the interaction effects (cf. Table 7, Panel B). Thus, hypotheses H3a and H3bhave to be neglected. Moreover, no significant direct effect of expertise could be found.

Finally, we checked whether there is a statistically significant concurrence between the user ratings and the expert evaluation. Therefore, the individual user ratings were aggregated and a quality ranking of the ideas was constructed according to the mean quality scores of the ideas. Then, correlation analysis was applied (cf. Table 8). In comparison to the expert rating the complex rating scale shows a strong, highly significant concurrence with $r = 0.62$ ($p < 0.01$). Neither the promote/demote nor the 5-star rating correlate with the expert rating. However, all rating scales show strong, very significant correlations among each other and in particular the aggregated idea rankings of the promote/demote scale are nearly identical ($r = 0.97$, $p < 0.001$).

Table 7. Moderated Regression Results					
Panel A. Moderating Effect of User Expertise on Rating Scale and Rating Accuracy					
Step	**Independent Variable**	**R^2**	**ΔR^2**	**Hypotheses**	**Supported**
1	Expertise	0.02	-		
2	Dummy 1	0.11**	0.09***		
	Dummy 2				
3	Expertise x Dummy1	0.12**	0.01	H3a	No
	Expertise x Dummy2				
Panel B. Moderating Effect of User Expertise on Rating Scale and Rating Satisfaction					
Step	**Independent Variable**	**R^2**	**ΔR^2**	**Hypotheses**	**Supported**
1	Expertise	0.03	-		
2	Dummy 1	0.08**	0.09**		
	Dummy 2				
3	Expertise x dummy1	0.10*	0.01	H3b	No
	Expertise x Dummy2				
N = 313, *** significant with p < 0.001, ** significant with p < 0.01, * significant with p < 0.05					

Table 8. Correlations of Expert Rating and Rating Scales			
	Expert Rating	**Promote/Demote Rating**	**5-star Rating**
Expert Rating			
Promote/Demote Rating	0.04		
5-star Rating	0.08	0.97***	
Complex Rating	0.62**	0.70***	0.68***
N = 24, *** significant with p < 0.001, ** significant with p < 0.01			

Summary and Discussion

Using questionnaire and system-captured experiment data, and an independent expert evaluation of idea quality, the proposed model was tested for relationships between the different rating scales and the resulting judgment accuracy and rating satisfaction. It was expected that the granularity of the rating scale would positively influence rating accuracy and positively influence users' satisfaction with their rating accuracy. Both of these hypotheses are supported (hypotheses H1 and H2). Moreover, it was expected that user expertise would have a moderating effect on the relationship between the rating scale and its rating accuracy and the users' rating satisfaction. These hypotheses in the model have not been supported (hypotheses H3a and H3b). We also tested for a direct effect of user expertise on rating accuracy but this has hypotheses has also not be supported.

Regarding the main condition of interest, rating accuracy, we reveal that the complex rating scale leads to a significantly higher rating accuracy than the promote/demote rating and the 5-star rating ($p < 0.001$). The measurements of both the individual users' rating accuracy measured by the fit-score (Figure 4) as well as the aggregated idea ranking agree in this finding (Table 8). Our results indicate that the highly popular promote/demote rating that is the current dominant design in innovation communities exhibits severe limitations in measuring idea quality. While it works well to identify top ideas as good and bottom ideas as bad it also produces the highest error (classifying top ideas as bad and bottom ideas as good). This results from a user bias of either rating very positively

(positivity bias, Tourangeau et al. 2000) or very negatively (e.g., for the 24 ideas a user would submit 20 promotes and only 4 demotes or vice versa). Thus, overall, the aggregated promote/demote rating is without insights regarding the measurement of idea quality and is near-random. This suggests the conclusion that this quick and easy decision making process fails. This can be explained in light of the cognitive decision process: while the rating scale has only little influence on the comprehension, information retrieval, and judgment phase, it has major influence on the reporting and response selection. Respondents failed to map their judgment on the two scale-endpoints of the binary rating scale. More granular rating scales offer more discretion for this mapping process thus leading to higher rating accuracy.

Another possible explanation could be that the different ratings scales do address different constructs associated with idea quality. The complex rating scale may represent a judgment of idea quality and the less granular scales an indication of idea popularity. However, idea quality and idea popularity do not necessarily have to be the same construct. Thus, they could activate different cognitive evaluation patterns in the decision process and leading to different results.

The judgment of rating accuracy, however, has to be seen in light of an optimal degree of granularity. The promote/demote rating, scores significantly lower regarding user satisfaction than the 5-star and the complex rating scale ($p < 0.01$) while the data shows no significant difference between the 5-star and complex rating scale. A possible explanation for this phenomenon is that the relationship between ratings scale granularity and rating satisfaction might rather have an inverted u-shape than being linear. The more granular a scale becomes the better the scale allows users to express their individual rating judgment more accurately which increases their rating satisfaction. According to Janis and Mann's (1977; 1982) conflict theory of decision making, the binary rating seems to elicit a major rating conflict resulting in high stress which cannot be resolved leading to low decision satisfaction and regret. More granular rating scales do expose idea raters to a more moderate level of stress. However, a too granular rating scale may reverse this effect as the accompanying rating effort rises. Thus, the 5-star rating seems to have an optimal degree of granularity in terms of rating satisfaction.

While our first two hypotheses are supported the hypotheses H3a and H3b regarding the moderating influence of user expertise have to be rejected. In addition to the moderating effect our analysis also found no direct effect of user expertise. Both these findings have important theoretical implications.

Regarding the analysis of a direct effect there is no significant difference between users with high and low expertise. This confirms the "wisdom of the crowds" theory that a larger group of people can perform decision tasks as good as experts irrespective of the knowledge of the individual. However, a key problem of the "wisdom of crowds" is the inability to distinguish between the "wisdom of the crowd" and "the mob that rules." The foundation of this problem lies in the improper usage of methods to delegate decision tasks to an anonymous group (Roman 2009). This hints at the importance of potential mediating effects as methods need to be designed to fit the target user group.

Regarding the analysis of a mediating effect there is no significant difference between users with high and low expertise regarding their use of the rating mechanism. Consequently, the "best" rating mechanism, i.e., the complex rating scale, performs best for all user groups, irrespective of their level of expertise. This leads to clear design recommendation that, with regards to rating accuracy, the complex rating scale should be used. Our experiment shows that in a well designed setting, a "crowd" can indeed perform similar to experts. The effective design of those rating mechanisms enhances the validity and reliability of resulting idea ratings and supports the selection of the best ideas for further refinement or implementation in a company's innovation process.

In summary, a combination of a web-based experiment, statistical analysis, and expert rating provides insights not possible with only one source of data and thus offers a fuller appreciation of the phenomena of online innovation communities. In particular, the test of moderating nature of user expertise allows deriving design guidelines regarding the IT system supporting online innovation communities. Overall, there is mutual support between the methodologies. The quantitative analysis of the "wisdom of crowds" hypothesis adds to our knowledge as to how a community can be used for tasks commonly performed by experts.

Conclusion

Theoretical and Methodological Contributions

Extending previous decision management research we offer insights into how different rating mechanisms for idea selection work within the context of online innovation communities. Contrary to the established practice of Internet-based rating which proposes that rating scales should be as simple as possible to avoid user drop-out, our research finds that very simple scales lead to near-random results. Consequently, more complex scales should be used, accepting higher drop-out rates but improving rating accuracy. Furthermore, prior research suggests that product knowledge is critical for reliable evaluation of innovation ideas. Our results contradict this. Our research finds no direct effect of user expertise on rating accuracy and also no mediating effect of user expertise on the influence of rating scale granularity on rating accuracy. This demonstrates that (1) in a well designed setting, a collective evaluation can match the performance of experts on a given evaluation task (direct effect), and (2) the more complex rating scales performs better for all user groups, irrespective of their expertise level (mediating effect).

Despite the widespread use of rating mechanisms in online innovation communities these popular tools have not yet been analyzed in depth. Our multi-method research is the first to offer reliable results comparing collective decision making with independent expert ratings, helping us to shed light into the question of how crowds can be engaged for certain tasks within complex decision making processes. Our research results in design guidelines favoring more complex rating mechanisms over simpler ones to improve, both decision quality, and user satisfaction.

Practical Implications

Effective and accurate design of mechanisms for collective decision making is critical to harness the wisdom of the crowds. If the design is ill-fitted to the desired task, outcomes can be misleading or simply wrong. Our research suggests that operators of popular innovation communities should re-consider their choice of using thumbs-up, thumbs-down ranking as it leads to, both near-random rating results, and low user satisfaction irrespective of user expertise. To improve user satisfaction and reliability of collective decision making operators of online innovation communities should opt for multi-attribute scales. While these scales might result in a lower number of submitted ratings due to higher drop-out rates the same psychometric attributes as with the promote/demote rating can be achieved with less ratings (King et al. 1983).

A possible design guideline can be given. An effective way of involving a community could be a combination of quality rating and popularity signaling. Instead of using promote/demote as a rating mechanism to judge idea quality, it should be used as a voting mechanism to signal popularity. To function as a signaling mechanism voting of other users should be visible. In a parallel approach, complex scales should be used to judge the actual idea quality. Here, ratings of other users should not be visible to avoid information cascades. To overcome issues with limited absorptive capacity by companies a combination of idea quality and idea popularity can then be used to decide which ideas to adopt based on popularity and actual idea quality.

Research Limitations and Future Directions

Some general shortcomings resulting from conducting a controlled experiment apply to our research. Through this research design users had no choice which ideas to rate as all ideas had to be rated. This might lead to a distortion of results regarding the promote/demote rating as this scale does not offer a neutral rating. Furthermore, following the "wisdom of the crowd" paradigm, the expert rating might be deficient as experts are more prone to a fixed mind-set than a broader community and thus might have overlooked certain aspects of some ideas.

Acknowledgements

Parts of this research have been funded through the GENIE project by the German Ministry of Research and Education (BMBF) under contract No. FKZ 01FM07027. This research also received funding from the German Federal Ministry of Economics and Technology (BMWi) under grant code 10MQ07024. The responsibility for the content of this publication lies with the authors.

References

Amabile, T.M. *Creativity in Context. Update to Social Psychology of Creativity*, (1 ed.) Westview Press, Oxford, UK, 1996.

Ang, S.H., and Low, S.Y.M. "Exploring the Dimensions of Ad Creativity," *Psychology & Marketing* (17:10) 2000, pp 835-854.

Bagozzi, R.P., and Yi, Y. "On the evaluation of Structural Equatation Models," *Journal of the Academy of Marketing Sciences* (16:1) 1988, pp 74-94.

Baron, R.M., and Kenny, D.A. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* (51:6) 1986, pp 1173-1182.

Benbasat, I., and Zmud, R. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties," *MIS Quarterly* (27:2) 2003, pp 183-194.

Besemer, S.P., and O'Quin, K. "Analyzing creative products: Refinement and test of judging tool," *Creativity Research Journal* (20:2) 1986, pp 115-126.

Besemer, S.P., and O'Quin, K. "Confirming the Three-Factor Creative Product Analysis Matrix Model in an American Sample," *Creativity Research Journal* (12:4) 1999, pp 287-296.

Biemer, P., and Lyberg, L.E. *Introduction to Survey Quality*, (1 ed.) John Wiley & Sons, Hoboken, NJ, USA, 2003.

Bloch, P., Ridgway, N., and Sherrell, D. "Extending the Concept of Shopping: An Investigation of Browsing Activity," *Journal of the Academy of Marketing Science* (17:1) 1989, pp 13-21.

Blohm, I., Bretschneider, U., Leimeister, J.M., and Krcmar, H. "Does Collaboration Among Participants Lead to Better Ideas in IT-Based Idea Competitions? An Ampirical Investigation," 43rd Hawaii International Conference on System Science (HICSS 43), Kauai, Hawai, 2010.

Boudreau, M.-C., Gefen, D., and Straub, D.W. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1) 2001, pp 1-16.

Browne, M.W., and Cudeck, R. "Alternative Ways of Assessing Model Fit," in: *Testing Structural Equation Models*, K.A. Bollen and J.S. Long (eds.), Sage, Newbury Park, CA, USA, 1993.

Bühner, M. *Einführung in die Test- und Fragebogenkonstruktion*, (2 ed.) Pearson Studium, München, Germany, 2008.

Cady, S.H., and Valentine, J. "Team Innovation and Perceptions of Consideration. What Difference does Diversity Make?," *Small Group Research* (30:6) 1999, pp 730-750.

Caroff, X., and Besançon, M. "Variability of Creativity Judgments," *Learning and Individual Differences* (18:4) 2008, pp 367-371.

Chesbrough, H. *Open Innovation: The New Imperative for Creating and Profiting from Technology* Harvard Business School Press, Boston, MA, USA, 2006.

Chesbrough, H., and Crowther, A.K. "Beyond High Tech: Early Adopters of Open Innovation in Other Industries," *R&D Management* (36:3) 2006, pp 229-236.

Christiaans, H.H.C.M. "Creativity as Design Criterion," *Creativity Research Journal* (14:1) 2002, pp 41-54.

Christian, L.M., and Dillman, D.A. "The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions," *Public Opinion Quarterly* (68:1) 2004, pp 57-80.

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

Cohen, J., Cohen, P., West, S.G., and Aiken, L.S. *Applied Multiple Regression / Correlation Analysis for the Behavorial Sciences*, (3 ed.) Erlbaum, Hillsdale, NJ, USA, 2003.

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

Collins, D. "Pretesting Survey Instruments: An Overview of Cognitive Methods," *Quality of Life Research* (12:3) 2003, pp 229-238.

Conrad, F.G., Couper, M.P., Tourangeau, R., and Peytchev, A. "Use and Non-use of Clarification Features in Web Surveys," *Journal of Official Statistics* (22:2) 2006, pp 245-269.

Cyr, D., Head, M., Larios, H., and Pan, B. "Exploring Human Images in Website Design: A Multi-Method Approach," *MIS Quarterly* (33:3) 2009, pp 539-566.

Dahlander, L., and Gann, D.M. "How Open is Open Innovation?," *Research Policy* (39:6) 2010, pp 699-709.

Dean, D.L., Hender, J.M., Rodgers, T.L., and Santanen, E.L. "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," *Journal of the Association for Information Systems* (7:10) 2006, pp 646-698.

Di Gangi, P.M., and Wasko, M. "Steal my Idea! Organizational Adoption of User Innovations from a User Innovation Community: A Case Study of Dell IdeaStorm," *Decision Support Systems* (48) 2009, pp 303-312.

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

Ebner, W., Leimeister, J.M., and Krcmar, H. "Community Engineering for Innovations -The Ideas Competition as a method to nurture a Virtual Community for Innovations," *R&D Management* (39:4) 2009, pp 342-356.

Enkel, E., Perez-Freije, J., and Gassmann, O. "Minimizing Market Risks Through Customer Integration in New Product Development: Learning from Bad Practice," *Creativity and Innovation Management* (14:4) 2005, pp 425--437.

Ferrando, P.J. "Testing the Equivalence Among Different Item Response Formats in Personality Measurement: A Structural Equation Modeling Approach," *Structural Equation Modeling*, (7:2) 2000, pp 271-286.

Finke, R.A., Ward, T.B., and Smith, S.M. *Creative Cognition. Theory, Research and Applications* MIT Press, Cambrigde, MA, USA, 1996.

Fishbein, M. "The Relationships Between Beliefs, Attitudes, and Behavior," in: *Cognitive Consistencey: Motivational Antecedents and Behavorial Consequents* S. Feldmann (ed.), Academic Press, New York, NY, USA, 1966.

Flynn, L.R., and Goldsmith, R.E. "A Short, Reliable Measure of Subjective Knowledge," *Journal of Business Research* (46:1) 1999, pp 57-66.

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

Franke, N., and Hienerth, C. "Prädikatoren der Qualität von Geschäftsideen: Eine empirische Analyse eines Online-Ideen-Forums," *Zeitschrift für Betriebswirtschaft Special Issue* (6:4) 2006, pp 47-68.

Franke, N., and Shah, S. "How Communities Support Innovative Activities: An Exploration of Assistance and Sharing Among End-Users," *Research Policy* (32:1) 2003, pp 157-178.

Frazier, P.A., Tix, A.P., and Barron, K.E. "Testing Moderator and Mediator Effects in Counseling Psychology Research," *Journal of Counseling Psychology* (51:1) 2004, pp 115-134.

Galletta, D.F., Henry, R., McCoy, S., and Polak, P. "Web Site Delays: How Tolerant are Users?," *Journal of the Association for Information Systems* (5:1) 2004, pp 1-28.

Ganassali, S. "The Influence of the Design of Web Survey Questionnaires on the Quality of Responses," *Survey Research Methods* (2:1) 2008, pp 21-32.

Gassmann, O. "Opening up the Innovation Process: Towards an Agenda," *R&D Management* (36:3) 2006, pp 223-228.

Goleman, D. *Emotional Intelligence. Why it can matter more than IQ*, (1 ed.) Bloomsbury, London, UK, 1996.

Griffin, M., Babin, B., and Attaway, J. "Anticipation of Injurious Consumption Outcomes and its Impact on Consumer Attributions of Blame," *Journal of the Academy of Marketing Science* (24:4) 1996, pp 314-327.

Janis, I.L., and Mann, L. *Decision Making. A Psychological Analysis of Conflct, Choice, and Commitment*, (1 ed.) The Free Press, New York, NY, USA, 1977.

Janis, I.L., and Mann, L. "A Theoretic Framework for Decision Counseling," in: *Counseling on personal decisions: Theory and helping on short-term helping relationships.*, I.L. Janis (ed.), Yale University Press, New Heaven, CT, USA, 1982.

Jeppesen, L.B., and Frederiksen, L. "Why do Users Contribute to Firm-Hosted User Communities? The Case of Computer-Controlled Music Instruments," *Organization Science* (17:1) 2006, pp 45-63.

Johnson-Laird, P.N. *Human and Machine Thinking*, (1. Aufl. ed.) Lawrence Erlbaum, Hillsdale, NJ, USA, 1993.

Jokisch, M. *Active Integration of Users into the Innovation Process of a Manufacturer. The BMW Customer Innovation Lab*, (1 ed.), München, Germany, 2007.

Keller, J.M. "Motivational Design of Instruction," in: *Instructional Design Theories and Models*, C.M. Reigeluth (ed.), Lawrence Erlbaum, Hillsdale, NJ, USA, 1983.

King, L.A., King, D.W., and Klockars, A.J. "Dichotomous and Multipoint Scales Using Bipolar Adjectives," *Applied Psychological Measurement* (7:2), April 1, 1983 1983, pp 173-180.

Kristensson, P., Gustafsson, A., and Archer, T. "Harnessing the Creative Potential Among Users," *The Journal of Product Innovation Management* (21:1) 2004, pp 4-14.

Lakhani, K., and Panetta, J. "The Principles of Distributed Innovation," *Innovations: Technology, Governance, Globalization* (2:3) 2007, pp 97-112.

LeDoux, J. *The Emotional Brain. The Mysterious Underpinning of Emotional Life*, (1 ed.) Weidenfeld & Nicolson, London, UK, 1998.

Leimeister, J.M. "Collective Intelligence," *Business & Information Systems Engineering* (2:4) 2010, pp 245-248.

Leimeister, J.M., Huber, M., Bretschneider, U., and Krcmar, H. "Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition " *Journal of Management Information Systems* (26:1) 2009, pp 197-224.

Lilien, G.L., Morrison, P.D., Searls, K., Sonnack, M., and von Hippel, E. "Performance Assessment of the Lead Iser Idea-Generation Process for New Product Development," *Management Science* (48:8) 2002, pp 1042–1059.

Lindgaard, G., and Dudek, C. "What is this Evasive Beast we Call User Satisfaction?," *Interacting witch Computers* (15:3) 2003, pp 429-452.

Lozano, L.M., García-Cueto, E., and Muñiz, J. "Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* (4:2) 2008, pp 73-79.

Lüthje, C. "Characteristics of Innovating Users in a Consumer Goods Field. An Empirical Study of Sport-Related Product Consumers," *Technovation* (24:9) 2004, pp 683-695.

MacCrimmon, K.R., and Wagner, C. "Stimulating Ideas Through Creative Software," *Management Science* (40:11) 1994, pp 1514-1532.

Malhotra, N.K. *Marketing Research. An Applied Orientation*, (5. Aufl. ed.) Pearson, Upper Saddle River, NJ, USA, 2007.

Matthing, J., Kristensson, P., Gustafsson, A., and Parasuraman, A. "Developping Succesful Technology-Based Services: The Issue of Identifying and Involving Innovative Users," *Journal of Services Marketing* (20:5) 2006, pp 288-297.

Mayer, R.E. "Fifty Years of Creativity Research," in: *Handbook of Creativity*, R.J. Sternberg (ed.), Cambrige University Press, Cambrige, MA, USA, 1999, pp. 449-460.

Nagasundaram, M., and Bostrom, R.P. "The Structuring of Creative Processes Using GSS: A Framework for Research," *Journal of Management Information Systems* (11:3) 1994, pp 87-114.

Niu, W., and Sternberg, R.J. "Cultural Influences on Artistic Creativity and its Evaluation," *International Journal of Psychology* (36:1) 2001, pp 225-241.

Ogawa, S., and Piller, F. "Reducing the Risks of New Product Development," *MIT Sloan Management Review* (47:2) 2006, pp 65-71.

Oliver, R.L., and Swan, J.E. "Consumer Perceptions of Interpersonal Equity and Satisfaction in Transactions: A Field Survey Approach," *Journal of Marketing* (53:2) 1989, pp 21-35.

Palvia, P., Leary, D., Mao, E., Midha, V., Pinjani, P., and Salam, A.F. "Research Methodologies in MIS: An Update," *Communications of the Association for Information Systems* (14) 2004, pp 526-542.

Piller, F.T., and Walcher, D. "Toolkits for Idea Competitions: A Novel Method to Integrate Users in New Product Development," *R&D Management* (36:3) 2006, pp 307-318.

Plucker , J.A., Beghetto, R.A., and Dow, G.T. "Why isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research," *Educational Psychologist* (39:2) 2004, pp 83-96.

Reinig, B.A., Briggs, R.O., and Nunamaker Jr, J.F. "On the Measurement of Ideation Quality," *Journal of Management Information Systems* (23:4), Spring2007 2007, pp 143-161.

Riedl, C., May, N., Finzen, J., Stathel, S., Kaufman, V., and Krcmar, H. "An Idea Ontology for Innovation Management," *International Journal on Semantic Web and Information Systems* (5:4) 2009, pp 1-18.

Rochford, L. "Generating and Screening New Product Ideas," *Industrial Marketing Management* (20:4) 1991, pp 287-296.

Roman, D. "Crowdsourcing and the Question of Expertise," *Communications of the ACM* (52:12) 2009, pp 12-12.

Runco, M.A., and Basadur, M. "Assessing Ideational and Evaluative Skills and Creative Styles and Attitudes," *Creativity and Innovation Management* (2:3) 1993, pp 166-173.

Runco, M.A., and Sakomoto, S.O. "Experimental Studies of Creativity," in: *Handbook of Creativity*, R.J. Sternberg (ed.), Cambridge University Press, Cambridge, MA, USA, 1999, pp. 62-92.

Runco, M.A., and Smith, W.R. "Interpersonal and Intrapersonal Evaluations of Creative Ideas," *Personality and Individual Differences* (13:3) 1992, pp 295-302.

Sainfort, F., and Booske, B. "Measuring Post-Decisions Satisfaction," *Medical Decision Making* (20:1) 2000, pp 51-61.

Schulz, C., and Wagner, S. "Outlaw Community Innovations," *International Journal of Innovation Management* (12:3) 2008, pp 399-418.

Schwarz, N. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation* Lawrence Erlbaum, Hillsdale, NJ, USA, 1996.

Shankar, V., Smith, A.K., and Rangaswamy, A. "Customer Satisfaction and Loyalty in Online and Offline Environments," *International Journal of Research in Marketing* (20:2) 2003, pp 153-175.

Sharma, R., Yetton, P., and Crawford, J. "Estimating the Effect of Common Method Variance: The Method-Method Pair Technique with an Illustration from TAM Research," *MIS Quarterly* (33:3) 2009, pp 473-490.

Small, R.V., and Venkatesh, M. "A Cognitive-Motivational Model of Decision Satisfaction," *Instructional Science* (28:1) 2000, pp 1-22.

Sniezek, J.A. "Groups Under Uncertainty: An Examination of Confidence in Group Decision Making," *Group Decision Making* (52:1) 1992, pp 124-154.

Spann, M., Ernst, H., Skiera, B., and Soll, J.H. "Identification of Lead Users for Consumer Products via Virtual Stock Markets," *Journal of Product Innovation Management* (26:3) 2009, pp 322-355.

Sudman, S., Bradburn, N.M., and Schwarz, N. *Thinking about Answers. The Application of Cognitive Processes to Survey Methodology*, (1 ed.) Jossey-Bass Publishers, San Francisco, CA, USA, 1996.

Tetlock, P.E. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press., Princeton, NJ, USA, 2006.

Torodova, G., and Durisin, B. "Absorptive Capacity: Valuing a Reconceptualization," *Academy of Management Review* (32:3) 2007, pp 774-786.

Toubia, O., and Flores, L. "Adaptive Idea Screening Using Consumers," *Marketing Science* (26:3) 2007, pp 342-360.

Tourangeau, R., Rips, L.J., and Rasinski, K. *The Psychology of Survey Response*, (1 ed.) Cambridge University Press, Cambridge, MA, USA, 2000.

Utterback, J. *Mastering the Dynamics of Innovation* Harvard Business School Press, Boston, MA, USA, 1996.

Voich, D. *Comparative Empirical Analysis of Cultural Values and Perceptions of Political Economy Issues* Praeger, Westport, CT, USA, 1995.

von Hippel, E. *The Sources of Innovation* Oxford University Press, New York, NY, USA, 1988.

von Hippel, E. ""Sticky Information" and the Locus of Problem Solving: Implications for Innovation," *Management Science* (40:4) 1994, pp 429-439.

von Hippel, E. *Democratizing Innovation* MIT Press, Cambridge, MA, USA, 2005.

Walcher, P.-D. *Der Ideenwettbewerb als Methode der aktiven Kundenintegration*, (1 ed.) Gabler, Wiesbaden, Germany, 2007.

West, J., and Lakhani, K. "Getting Clear About Communities in Open Innovation," *Industry and Innovation* (15:2) 2008, pp 223-231.

West, S.G., Aiken, L.S., and Krull, J.L. "Experimental personality designs: Analyzing categorical by continuous variable interactions," *Journal of Personality* (64:1) 1996, pp 1-49.

Wheaton, B., Muthén, B., Alwin, D.F., and Summers, G.F. "Assessing Reliability and Stability in Panel Models," in: *Sociological Methodology*, D.R. Heise (ed.), Jossey-Bass, San Francisco, CA, USA, 1977.

Zajonc, R. "Feeling and Thinking: Preferences Need no References," *American Psychologist* (35:2) 1980, pp 151-175.

Appendix A – Experiment Web Sites

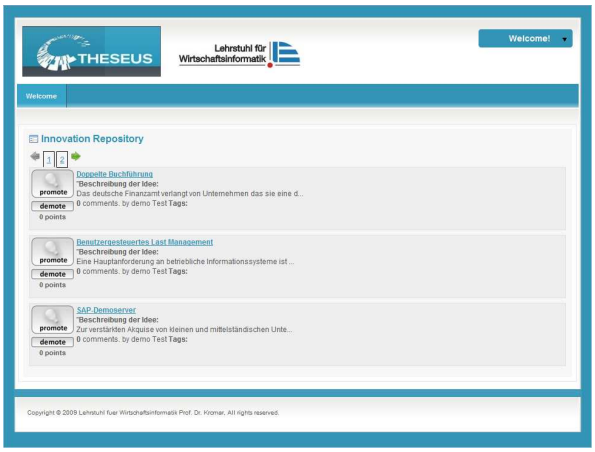

Figure A1. Promote/Demote Rating Scale

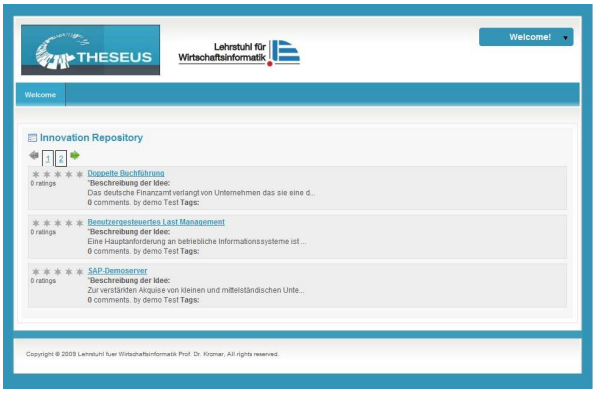

Figure A2. 5-star Rating Scale

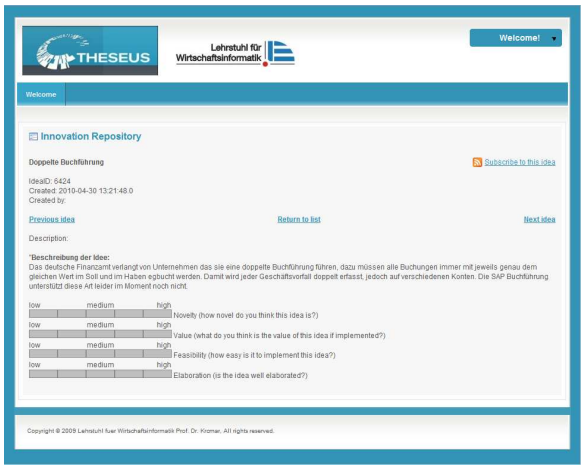

Figure A3. Complex Rating Scale: Complex rating scale on the details page of an idea: four 5-point scales for (1) novelty, (2) value, (3) feasibility, and (4) elaboration ranging from "low" to "high".

Appendix B – Items Measuring User Satisfaction and Expertise

User Expertise	
EXP1	I have a great deal of skill in using SAP software.
EXP2	I know pretty much about SAP software and the company in general.
EXP3	The features of SAP software are well-known to me.
EXP4	I know how to operate SAP software.
Rating satisfaction	
SAT1	The idea rating was user friendly.
SAT2	I am satisfied with my ratings.
SAT3	The scale did not reflect my true perception of idea quality (reverse coded).
SAT4	Rating the ideas met my expectations.