

Please quote as: Blohm, I.; Riedl, C.; Leimeister, J. M. & Krcmar, H. (2011): Idea Evaluation Mechanisms for Collective Intelligence in Open Innovation Communities: Do Traders outperform Raters?. In: 32. International Conference on Information Systems, Shanghai, China.

IDEA EVALUATION MECHANISMS FOR COLLECTIVE INTELLIGENCE IN OPEN INNOVATION COMMUNITIES: DO TRADERS OUTPERFORM RATERS?

Completed Research Paper

Ivo Blohm

Technische Universität München
Chair for Information Systems
Boltzmannstr. 3
85748 Garching b. München
ivo.blohm@in.tum.de

Christoph Riedl

Harvard University
Institute for Quantitative Social
Science
1793 Cambridge St.
Cambridge, MA 02138
criedl@iq.harvard.edu

Jan Marco Leimeister

Universität Kassel
Chair for Information Systems
Nora-Platiel-Str. 4
34127 Kassel
leimeister@uni-kassel.de

Helmut Krcmar

Technische Universität München
Chair for Information Systems
Boltzmannstr. 3
85748 Garching b. München
krcmar@in.tum.de

Abstract

The increasing popularity of open innovation approaches has led to the rise of various open innovation communities on the Internet which might contain several thousand 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 mechanisms supporting the evaluation of these ideas. In this paper, we focus on the evaluation of such mechanisms for collective idea evaluation. Applying a multi-method approach, we compare six different configurations of a prediction market with a multi-criteria rating scale that performed best in previous research. We combine a web-based experiment with 448 participants, data from a participant survey, and an independent expert jury. Based on cognitive load theory, we explain why a multi-criteria rating scale outperforms prediction markets in terms of evaluation accuracy and evaluation satisfaction. This study contributes to theory building in the emerging field of collective intelligence.

Keywords: open innovation, collective intelligence, rating scales, prediction markets, idea evaluation, communities, crowdsourcing, cognitive load theory, information aggregation, web 2.0

Introduction

Over the last decade, the concept of open innovation (OI) has been firmly established in research and practice (e.g., Enkel et al. 2005; Gassmann 2006; Lakhani and Panetta 2007). Open Innovation refers to an innovation regime in which an organization opens its innovation processes for both an inflow and an outflow of ideas to external parties (Chesbrough 2006). Today, companies can no longer rely solely on internally developed innovations but systematically integrate other sources of innovation in order to improve their innovativeness. Customers, in particular, are seen as one of the biggest resources for ideas for innovations (Bogers et al. 2010; Enkel et al. 2005; von Hippel 2005). The positive impact of customer integration on company success has been demonstrated in various studies (e.g., Enkel et al. 2005; Gassmann 2006; Ogawa and Piller 2006; von Hippel 2005; West and Lakhani 2008). One key method of integrating customers into innovation development are OI communities. Innovation communities invite external actors, in particular end-users, to freely reveal innovative ideas (von Hippel 2005). Through these communities, members contribute their ideas to be reviewed, discussed, and rated by the user community (Blohm et al. 2011a; Ebner et al. 2009; Franke and Shah 2003; Riedl et al. 2009). Prominent examples are Dell IdeaStorm or MyStarbucksIdea, both comprising several thousand users and user-generated ideas. While these OI communities provide access to the knowledge of many customers and thus to a large amount of need and solution information (von Hippel 2005), one core challenge arises for companies: How to select and filter the most relevant information from those communities? Constraints of time, budget, cognitive resources, and organizational structures limit the absorptive capacity of community owners, so that only a fraction of submitted ideas can be implemented (Blohm et al. 2011b; Cohen and Levinthal 1990; Di Gangi and Wasko 2009). Mechanisms for community-based idea evaluation may facilitate the process of identifying the 'best' ideas (Berg-Jensen et al. 2010). As the evaluation of external information is a crucial facet of absorptive capacity (Torodova and Durisin 2007), the application of these mechanisms may enhance the incorporation of ideas. Applying Shannon's (1948) information theory the community evaluation can act as filter between the community and the company, thus enhancing the quality of the knowledge transmission between the community as sender and the company as receiver of ideas. With appropriately designed filters, high quality ideas can be identified more easily, as the 'noise' reflecting low quality ideas can be sorted out, and thus companies only have to cope with a smaller number of ideas. In the context of this work we define a mechanism's evaluation accuracy as its ability to identify the best ideas from the viewpoint of the adopting organization and a to achieve high 'fit' with the organization's own assessment, which is commonly performed by in-house experts (Cohen and Levinthal 1990; Lane et al. 2006).

Mechanisms for utilizing the collective intelligence of crowds (Leimeister 2010) for evaluation purposes are still rare (Bonabeau 2009), even though they can serve as decision support systems for companies (Berg and Rietz 2003). Evaluating the quality of ideas in OI communities can be considered as a collective judgment task (Zigurs and Buckland 1998) and in current studies, rating scales (e.g., Di Gangi and Wasko 2009; Riedl et al. 2010) and prediction markets (e.g., Dahan et al. 2010; LaComb et al. 2007; Soukhoroukova et al. 2011) have been researched as two major mechanisms for collective idea evaluation. Regarding the use of rating scales for idea evaluation, Riedl et al. (2010) suggest that more elaborate scales involving multiple attributes perform better than simple scales, despite their popularity. Besides, prediction markets have been found to be very promising for the evaluation of innovation ideas (Bothos et al. 2009; Soukhoroukova et al. 2011). However, the design of trading mechanisms is affecting the behavior of users, their satisfaction, and overall market efficiency (Chen et al. 2010; Jian and Sami 2010) and it is unclear how these markets should be set-up for idea evaluation. As participation in OI communities fluctuates, appropriate evaluation mechanisms should not only be accurate. They should also create high satisfaction among users, as users will simply stop evaluating ideas if this is perceived as dissatisfying. Eliciting high participation and enabling participants to make valuable contributions are important challenges in the design of social interaction systems for collective idea evaluation. Thus, user perceptions of such mechanisms have to be incorporated into the overall design decisions since they are pivotal antecedents of a wise crowd (Graefe 2009). In addition to uncertainty regarding the design of the individual mechanism, the question arises how these two different concepts stack up against each other as they lack empirical evaluation of their relative performance (Chen et al. 2005; Graefe 2009).

To answer the question of how to design the mechanism of a prediction market and to perform a relative performance comparison between prediction markets and rating scales as idea evaluation mechanisms, we followed a two-stage experimental design with multiple treatment groups (Shadish et al. 2002) (cf. Figure 1). In stage I, we compare six different prediction markets using Hanson's (2003; 2007) market maker to identify robust configurations of such a mechanism. In stage II, we compared the best performing prediction market from stage I against a multi-criteria rating scale that was found to be most appropriate for evaluating innovation ideas in previous research (Riedl et al. 2010). Both stages of the experiment employed a between subject design with random assignment of 448 participants to seven treatment conditions (six markets and one rating scale). In both stages, independent expert evaluations served as baseline for the performance comparison. To ensure that the comparison 'prediction market vs. rating scale' (the main focus of this study) is robust, we need to ensure that both mechanisms, individually, follow the best possible design. While we are able to build on prior research regarding the design of the rating scale, no clear best-of-breed recommendation is available for prediction markets. Hence, we follow this two-stage experimental design in which we first analyze different configurations of prediction markets before comparing it to the best-performing rating scale suggested by prior research.

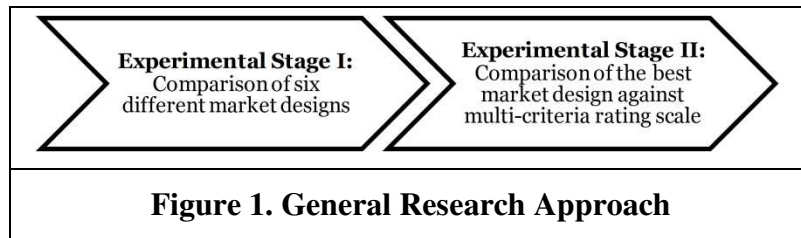


Figure 1. General Research Approach

In summary, this study has the following goals:

1. From a theoretical perspective, we create and test a model to analyze the performance of two major concepts for eliciting group judgments: a prediction market and a rating scale. Furthermore, we evaluate six different market designs for their relative performance. Thus, our paper provides a first experimental comparison of these two mechanisms for group judgments.
2. From a methodological perspective, we apply three different research and data collection methods to analyze the effectiveness of two different idea evaluation mechanisms that are commonly used in OI communities. Triangulating a web experiment, a questionnaire, and independent expert evaluation of idea quality, we are able to increase the robustness of our results.
3. From a practical perspective, our research provides actionable design guidelines for community-based idea evaluation mechanisms in OI communities. Following these design recommendations, community evaluations in OI communities should be improved.

The paper is structured as follows: After the following short review of the theoretical basis, we present our research model and hypotheses. We then present our research design including a detailed description of the experimental set-up. The next section then analyzes six different prediction market configurations. This is followed by the main analysis comparing the best-performing prediction market against a multi-criteria scale and the test of our hypotheses. Finally, the conclusion discusses contributions, limitations, and opportunities for future research.

Theoretical Background

Idea Quality

Community-based idea evaluation mechanisms aim to filter out the best ideas among a pool of submissions. Rating scales and prediction markets function as alternative ways of assessing idea quality. In this context, idea quality is a complex construct consisting of four distinct dimensions: novelty, feasibility, relevance and elaboration (Blohm et al. 2011a). Originality is the most important facet of novelty and refers not only to an idea being unique or rare, but also to being surprising, imaginative, or uncommon (Dean et al. 2006). From a new product development perspective, the novelty of an idea refers to its innovativeness. Novelty alone is not enough for an idea to be useful. Ideas must also solve a tangible,

and vital problem (Amabile 1996; Dean et al. 2006). Usefulness refers to an idea's value or relevance (Kristensson et al. 2004). In the scope of OI communities, this would be the financial potential and the customer benefit that an idea endows (Cady and Valentine 1999). An idea's feasibility is another vital dimension of idea quality, as it captures the ease with which the idea can be transformed into a commercial product (Cady and Valentine 1999). Another trait of idea quality is elaboration, which is the extent to which idea are complete, detailed, and understandable (Dean et al. 2006).

Idea Evaluation Mechanisms

Rating Scales

Applying rating scales, users choose among a finite set of competing alternatives by evaluating a defined set of criteria. By assigning numerical values to the given criteria, rating scales strive for identifying an alternative that is closest to an a priori defined optimum (Limayem and DeSanctis 2000). By means of different weighting and aggregation algorithms as used in Multi-Criteria Decision Making (Triantaphyllou 2000) or opinion pools (Chen et al. 2005), individual ratings can be aggregated to group decisions. Apart from their original domain of social science (Christian et al. 2007; Couper et al. 2007), the psychometric properties, perception, and usage of rating scales for online use have been researched in the fields of human-computer interaction (Knapp and Kirk 2003; van Schaik and Ling 2007), e-commerce (Cosley et al. 2003; Winkelmann et al. 2009), knowledge management (Poston and Speier 2005), and OI (Berg-Jensen et al. 2010; Di Gangi and Wasko 2009; Riedl et al. 2010).

Prediction Markets

Prediction markets are virtual market places on which participants trade contracts that are bound to the occurrence of a future event and whose purpose is to collect, aggregate, and evaluate dispersed information (Arrow et al. 2008; Spann and Skiera 2003; Wolfers and Zitzewitz 2004). The theoretical foundation of prediction markets is the efficient market hypothesis. According to Hayek (1945), market prices are the most efficient instrument to aggregate asymmetrically dispersed information. Thus, market prices in efficient markets can be used for forecasting as they reflect all available information (Fama 1970; 1991). In prediction markets, participants buy contracts that have a certain payoff (e.g., \$100) if a future event occurs. In the case that this event does not occur, contract holders receive nothing. Thus, the market price reflects the probability that an event occurs, and traders can make profits if they correctly predict the event's occurrence. Prediction markets have successfully been used in the domains of politics (Forsythe et al. 1999), sports (Chen et al. 2005; Servan-Schreiber et al. 2004), and economics (Spann and Skiera 2003). Researchers also applied the concept to the evaluation of new product ideas (Bothos et al. 2009; LaComb et al. 2007; Soukhoroukova et al. 2011), new product concepts (Dahan et al. 2010) and early stage technologies (Chen et al. 2009-10). A major concern of prediction markets are 'thin markets' in which information aggregation is ineffective due to insufficient traders (Hanson 2003; Healy et al. 2010). Automatic market makers overcome this problem with algorithms that adjust prices based on the transactions of the traders (Boer et al. 2007; Das 2005). They give instant feedback to traders, as trades can be performed at any time without having to wait for a second trader as a counterparty (Berg and Proebsting 2009; Pennock and Sami 2008). Hanson's (2003; 2007) Logarithmic Market Scoring Rules (LMSR) maker is currently the most applied market maker (Jian and Sami 2010; Othman and Sandholm 2010) and has been applied in politics and sports (Slamka et al. 2011), general economic forecasts (Mizuyama and Komatsu 2010; Othman and Sandholm 2010), and idea markets evaluation. Moreover, it is feasible to handle large-scale experiments (Gaspoz and Pigneur 2008; Othman and Sandholm 2010).

Comparison of Rating Scales and Prediction Markets

Rating scales and prediction markets are fundamentally different in their approaches to evaluate idea quality. First, using a rating scale, an absolute assessment of a rating object is created which is set against this scale. This rating has a meaningful interpretation on its own. A prediction market, on the other hand, creates a relative comparison of all rating objects against themselves. Consequently, for a meaningful interpretation all rating objects that were part of the relative comparison must be known. A related distinction can be made regarding the social properties of the two mechanisms. While a rating scale can be operated by a single user in isolation, prediction markets are per definition social mechanisms that require several traders to perform trades. Furthermore, rating scales can be operated in a single, one-off fashion whereas in a prediction market, repeated trading and involvement over time would be required to

aggregate information. A final distinction can be made regarding the final output of the rating mechanisms. A rating scale produces individual ratings by each user for each rating object. These individual ratings then have to be aggregated using some kind of aggregation mechanism (e.g., aggregation by mean) to arrive at an overall ranking. A prediction market, on the other hand, does not produce individual ratings but produces only the overall ranking of all rating objects.

Cognitive Load Theory

Research on computer-human interaction suggests that the success of task completion is a function of information presentation and the users' cognitive abilities (Nielsen 1994; Shneiderman and Plaisant 2004; Stewart and Travis 2003). In this regard, cognitive load theory (Sweller 1988) assumes that the human cognitive architecture utilizes a short-term working, and a long-term storage memory, processing information and tasks. Human working memory processes all conscious cognitive tasks. As the number of interacting elements that can be handled by the working memory is very limited (Baddeley 1986; Cowan 2005; Miller 1956), human cognition expands on the long-term memory. In long-term memory, cognitive structures – schemas – are created which incorporate multiple elements of the working memory and assign a specific function to them. Schema usage frees cognitive resources in the working memory and thus enhances its capacity. Cognitive overload arises if a task's information processing demands exceed the processing capacity of the cognitive system of the task fulfiller (Mayer and Moreno 2003). Cognitive overload inhibits human information processing and lowers decision quality (Hwang and Lin 1999) as it may lead to a state of 'bounded' rationality in which humans cannot consider and integrate all available information into their judgment (March 1978). Cognitive load theory was originally developed in the domain of student problem solving (Sweller 1988) and was applied to various IS related contexts such as the design of e-learning (e.g. Brünken et al. 2003; Mayer and Moreno 2003), e-commerce (Schmutz et al. 2009), search applications (Gwizdka 2010), multi-modal user interfaces (Berthold and Jameson 1999; Leung et al. 2007), or usability research in general (Chalmers 2003; DeStefano and LeFevre 2007).

Hypotheses and Model Development

Advocates of prediction markets argue that prediction markets are a powerful forecasting and evaluation method, as they are very effective in aggregating dispersed information from multiple respondents. However, we believe that prediction markets are inappropriate for idea evaluation in OI communities as their inherent complexity may overload users cognitively hampering the mechanism's effectiveness.

The evaluation of the quality of user-generated innovation ideas is a complex task which may not have a true solution and is thus inducing a high cognitive load (van Merriënboer et al. 2003). Cognitive load is additive and consequently ill-designed mechanisms may add additional load to a users working memory. Paas et al. (2003) postulate three sources of cognitive load: (1) intrinsic cognitive load, (2) extraneous cognitive load, and (3) germane cognitive load. (1) Intrinsic cognitive load refers to task-inherent complexity and keeping a mental representation of the task in the working memory over a period of time (Mayer and Moreno 2003). Whereas rating scales are easy and intuitive for most users (Winkelmann et al. 2009), prediction markets provide a considerably higher complexity (Graefe 2009; Kamp and Koen 2009). Understanding and incorporating the market logic representing idea quality requires participants to process more interacting elements, and induces a higher intrinsic cognitive load than using rating scales. (2) Whereas intrinsic cognitive load directly refers to the complexity of the task, extraneous cognitive load refers to its presentation and arises from changes in information architecture, visual complexity, and additional media use (Mayer and Moreno 2003) as well as rising information quantity (information overload) (Eppler and Mengis 2004; Kirsh 2000). Prediction markets present a broad array of additional financial information, such as graphical representations of market prices as developments in order to support decision-making of its users. Moreover, prices and relating information constantly change due to the transactions of traders, thus reflecting a higher extraneous cognitive load in comparison to rating scales. (3) Finally, germane cognitive load refers to learning and the ease with which schemas in the long-term memory can be created (Paas et al. 2003). Most potential users of OI communities will have used rating scales before, e.g., in the scope of questionnaires they have completed or other web 2.0 applications they may have used (Winkelmann et al. 2009). By contrast, prediction markets are relatively sparsely used suggesting that only few users have used them before. Whereas existing schemas can easily be adapted for rating scales, most participants will have to create new schemas in order to use prediction

markets. Thus, prediction markets are very likely to induce a higher germane cognitive load, and positive learning effects of schema usage will occur at a slower pace. Summing up, we believe that rating scales impose a lower intrinsic, extraneous, and germane cognitive load; thus, we assume:

H1: The evaluation mechanism used influences evaluation accuracy such that the mechanism 'rating scale' leads to higher accuracy than the mechanism 'prediction market.'

Generally, satisfaction has an evaluative focus, reflecting how favorable or unfavorable a person is toward a specific alternative (Fishbein 1966). Post-decision satisfaction arises immediately after the decision (Sainfort and Booske 2000) and according to Janis and Mann's (1977; 1982) conflict theory of decision making, decision satisfaction is heavily influenced by decisional stress. 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 is generally termed 'feeling' accompanies all cognitions. Thus, judgments of objective properties, such as evaluating idea quality, are often influenced by affective reactions (LeDoux 1998; Zajonc 1980). Existing research has found that cognitive overload is negatively associated with decision satisfaction (Grise and Gallupe 2000). In this regard, overload creates uncertainty (Botti and Iyengar 2006), cognitive dissonance (Malhotra 1982), and feelings of frustration (Farhoomand and Drury 2002) that are all facets of high decisional stress imposed by unresolved decision conflicts (Janis and Mann 1977; 1982). Thus, the higher cognitive load induced by prediction markets in comparison to rating scales is more likely to shift users into a state of 'hyper-vigilance,' a mismatch between the user's expectations of evaluating the ideas accurately and the perceived accuracy of the evaluation. Most rating scale users will have used rating scales before so that there will be only little uncertainty regarding their correct application. Rating scale users can focus their full attention to the evaluation of the ideas reducing decisional stress whereas prediction market users will have to employ cognitive resources on the appropriate mechanism application as well. Moreover, rating scales employ a direct evaluation of ideas during which users can assign low values to bad and high values to good ideas in a straightforward manner. In contrast, prediction market users have to consider various constraints when making their decision that stem from the indirect market logic, e.g., the available money in their user account. These constraints may prevent users from making their quality judgments as they would without leading to higher decisional stress. Thus, dissatisfaction or even frustration that have risen from the feeling of not having been able to express its own judgment are much likelier to occur on prediction markets. Summing up, we assume that prediction markets lead to a lower degree of evaluation satisfaction as they elicit more stressful conflict situations than rating scales that users cannot adequately cope with, and we propose:

H2: The evaluation mechanism used influences evaluation satisfaction such that the mechanism 'rating scale' leads to higher satisfaction than the mechanism 'prediction market.'

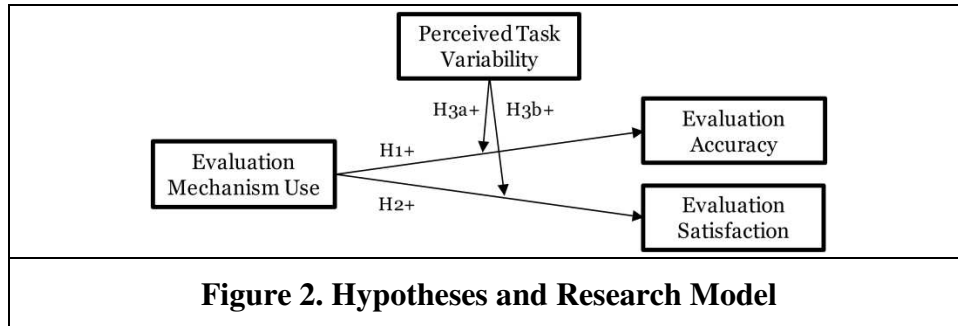
Generally, cognitive load rises when task complexity increases (Paas et al. 2003; Sweller 1988). Task complexity is highly determined by task variability (Perrow 1967) which is the number of exceptional cases encountered solving a task, and refers to the frequency of novel and unexpected results (Daft and Macintosh 1981). In comparison to rating scales, prediction markets induce a higher extraneous cognitive load as they display a broad array of additional financial market information that is changing with every user transaction. Due to these dynamics of constantly changing prices and market conditions, evaluating ideas with prediction markets instead of rating scales is a considerably more variable task. However, the perception of task characteristics, such as task variability, is highly subjective (O'Reilly et al. 1980), and is heavily dependent on already existing cognitive structures (Haerem and Rau 2007). Cognitive load should thus be higher for individuals perceiving a task highly variable than it is for individuals with low perceived task variability (PTV). Consequently, participants perceiving high task variability are more endangered to face cognitive overload which hampers their evaluation accuracy. Given the higher cognitive load of prediction markets in comparison to rating scales users with high PTV are more endangered of experiencing cognitive overload when using a prediction market instead of a rating scale. Thus, users perceiving the task of evaluating idea quality highly variable should have a higher evaluation accuracy using rating scales than using prediction markets. Contrarily, the risk of cognitive overload is smaller for users with low PTV. Accordingly, the difference in evaluation accuracy between users of prediction markets and rating scales should be smaller. Thus, we assume:

H3a: Perceived task variability positively moderates the effect of the evaluation mechanism used on evaluation accuracy such that the difference in evaluation accuracy between the mechanism ‘rating scale’ and the mechanism ‘prediction market’ will be higher for high perceptions of task variability and smaller for low perceptions of task variability.

Research has found task variability to be negatively associated with user satisfaction (Speier et al. 2003). Generally, a task becomes less predictable as its variability increases. Decision makers can rely less extensively on existing cognitive structures, and have to invest more cognitive resources in understanding and sense making (Karimi et al. 2004). As a consequence, decisional stress increases that should generally be higher for individuals with high PTV than for individuals with low PTV. Thus, users with high PTV are more likely to undergo a state of hyper-vigilance which decreases evaluation satisfaction when the assess idea quality by means of a prediction market instead of rating scale. On the flipside, evaluation satisfaction among users with low PTV should converge on both mechanisms as these users have enough free cognitive resource in order to prevent cognitive overload and resulting dissatisfaction, Thus, we propose:

H3b: Perceived task variability positively moderates the effect of the evaluation mechanism used on evaluation satisfaction such that the difference in evaluation satisfaction between the mechanism ‘rating scale’ and the mechanism ‘prediction market’ will be higher for high perceptions of task variability and smaller for low perceptions of task variability.

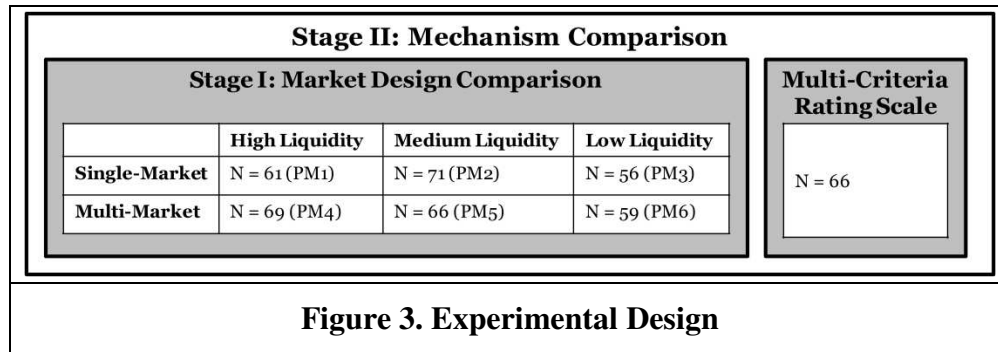
By consolidating all four hypotheses the research model in Figure 2 emerges.



Research Design

The two-stage experiment was designed as a web-based experiment using a standard innovation portal for OI communities developed by the authors. Standard features, such as idea submissions, commenting, and sorting were disabled, and only the evaluation mechanisms were activated. Apart from these mechanisms, all portals were identical (cf. Appendix A). The portal consisted of a summary page containing the ideas to be evaluated, a portfolio page containing the ideas that a user has evaluated, and a FAQ explaining the experimental task as well as the mechanism’s way of functioning. For the prediction market, the portfolio page contained additional financial information, such as transaction prices, liquid funds, and a graph representing a trader’s overall portfolio value. The system provided visual feedback for a successful idea evaluation (e.g., highlighting rating scale buttons or updating price graphs) making user interaction as easy as possible. Subjects participated with their own computers. Before starting the experiment, we confirmed whether all common web browsers displayed the innovation portal correctly. As a web experiment closely reflects the actual usage scenarios of OI communities, high external validity of our results can be assumed. Participants could evaluate the ideas in their natural environment and could allocate as much time to complete the task as desired. Furthermore, the internal validity of our results was enhanced by analyzing the log files of the portals. By doing so, inappropriate user behavior, such as a quick random evaluation of the ideas, could be identified in order to exclude these users from the analysis.

Figure 3 shows the overall research design of our experiment. In stage I we identify a robust configuration of the prediction market in terms of evaluation accuracy and satisfaction, which we then compare with a multi-criteria scale in stage II.



Stage I was designed as a randomized 2x3 between subject factorial experiment (cf. Figure 3 ‘Stage I’). The first factor represented the liquidity of the market maker, which we varied from low to high. The liquidity of the market maker can be defined as the degree to which prices change based on a single transaction. The second factor was market design, where we implemented ‘single-market’ and ‘multi-market’ designs. The single-market design handled all contracts for all ideas on one single market, whereas contracts for each idea are traded on separate markets on ‘multi-markets.’ A detailed description of these factors can be found in the section ‘Prediction Market Configurations’ below. Each treatment group consisted of at least 56 participants exceeding the recommendation of 30 participants for efficient prediction markets (Christiansen 2007). The forecasting goal of our markets was set to identify the best five ideas. Intensive pretesting revealed that the participants perceived the task of identifying the best five ideas as considerably easier than identifying the best idea. On all markets participants received a seed capital of 5,000 virtual currency units. Participants received a payoff of 100 virtual currency units for each idea-contract in their portfolio that were correctly classified and 0 for incorrect classifications.

In experimental stage II, we applied a 2x1 between subject factorial design which compared the best-performing prediction market from the first stage with a multi-criteria rating scale representing the best-performing rating scale configuration, as identified by Riedl et al. (2010) (cf. Figure 3 ‘Stage II’). In order to avoid a position bias in the rating scale treatment, ideas were presented in a randomized order (Malhotra 2007). Furthermore, rating scale users were able to update their ratings, so that it was assured that every user could rate every idea only once. In order to avoid information cascades in the rating scale treatment (Easley and Kleinberg 2010) leading to a rating bias derived from other participants’ ratings, the rating information of other participants was not visible. The multi-criteria rating scale used the following four criteria: novelty, value, feasibility, and elaboration (Riedl et al. 2010).

In both stages of the experiment, participants of the sample population were randomly assigned to one of the evaluation mechanism treatments (random sampling without replacement; between subject design). Based on the random assignment, we invited the participants via a personalized email that included a link with the respective system URL and the online questionnaire, as well as an exhaustive description of the experimental task. Additionally, we provided all participants with a unique activation code that was necessary upon registration on the innovation portal in order to prevent cross-contamination effects and manipulations through the creation of multiple user accounts. Manipulation is especially a problem in prediction markets where participants can easily transfer money from one account to another by trading against themselves (Blume et al. 2010). In both stages, the participants completed the idea evaluation task distributed over the experiment duration of three weeks in November 2010.

Idea Sample

The ideas to be evaluated in the experiment comprised of a title and a description. The ideas were taken from a German real-world OI community of the software producer SAP. In this community, SAP users are invited to submit ideas to improve the SAP software or to develop radical innovations in the scope of the SAP software. Currently, it consists of 285 users who have submitted 208 ideas varying in length between a half and full A4 page. An independent panel of experts evaluated all ideas. Among all ideas, idea quality is normally distributed (Kolmogorov-Smirnov Z-score: 0.56, p = 0.91). Since conducting an experiment with all ideas implied a substantial workload for participants a stratified sample of 24 ideas was drawn. This sample comprised 8 ideas each with high, medium, and low quality. The sample size was considered

sufficient, as 20 to 30 ideas are generally used to measure the variance of creativity ratings of laypersons in creativity research (Caroff and Besançon 2008; Runco and Basadur 1993; Runco and Smith 1992).

Participants

Users of topic related OI portals can be seen as the target population of our study and OI communities in general. Users of OI communities are predominantly male, young, and well educated (Franke and Shah 2003; Jeppesen and Frederiksen 2006; Jokisch 2007; Kristensson et al. 2004; Schulz and Wagner 2008). 486 participants took part in the experiment and 448 were included in the analysis. Our sample consisted of undergraduate and graduate students from two information systems courses related to SAP, as well as research assistants from the same field at a large German university. In order to motivate the participants, we offered homework credit points for participating students and drew two mp3 players for the participants with the highest concurrence with the expert jury (similar to Slamka et al. 2011). This payout scheme corresponds to a rank-order tournament that was found to enhance accuracy of prediction markets (Luckner and Weinhardt 2007), and we assumed that this would have a similar effect for rating scales. 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 required knowledge of SAP software systems to judge idea quality. It can also be argued that IS students are suitable participants, as they represent actual users of OI communities. On a general level, Voich (1995) found the values and beliefs of students to be representative in a variety of occupations. We applied Multivariate Analysis of Variance in order to check random assignment of participants, and found no differences regarding age, gender, and educational level in the different treatment groups. There were no significant differences between students and research assistants. 74.9% of our subjects were male, 11.5% had a master degree, 24.9% a bachelor degree and 58.3% finished high school. Mean age of participants was 22.54 years.

Data Sources

We combined three research and data collection methods to identify an effective prediction market, and to compare it with a multi-criteria rating scale: (1) a web experiment reflecting users' idea evaluations, (2) a quantitative survey of participants, and (3) an independent expert rating of idea quality. This triangulation allows detailed insights into the complex interaction of user behavior, satisfaction, and IT artifacts. Furthermore, various researchers advocate the use of multiple methods to gain more robust results overcoming common method bias (Boudreau et al. 2001; Cyr et al. 2009; Sharma et al. 2009).

Experiment Idea Evaluations

Initially, 486 participants took part in the experiment. Subjects that did not complete the survey and/or performed the task in less than five minutes were removed from the analysis. The remaining 448 participants performed 13,678 transactions in the prediction market treatments and 5,752 individual ratings in the rating scale treatment. The median time it took the users to participate was 78 minutes and 17 seconds.

Questionnaire

Data on the participants' PTV and evaluation satisfaction was collected through an online questionnaire after the experiment. For measuring PTV, we adapted the scales of Whitey et al. (1983) that were formerly used in the context of comparing task perceptions of novices (Haerem and Rau 2007). For measuring evaluation satisfaction, we adapted the scales of Riedl et al. (2010) previously used for measuring evaluation satisfaction of rating scales for idea evaluation. All items were measured on a 5-point Likert scale. The entire survey was pretested with a small sample of ten participants, reflecting the different groups of participants. 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

In practice, companies usually evaluate innovation ideas with a small group of experts (Ferioli et al. 2010; Ozer 2005; Rochford 1991; Toubia and Flores 2007; Urban and Hauser 1993). Accordingly experts are generally used in order to identify the most promising ideas in OI communities (Berg-Jensen et al. 2010; Bretschneider 2011). In this regard, the expert evaluations provide a proxy measure for actual idea

quality, which is not observable. Thus, we compared the experiment participants' evaluations with an independent expert evaluation in order to assess the accuracy of the different mechanisms. The ideas from the OI community were evaluated by a jury using the consensual assessment technique (Amabile 1996). This technique is derived from creativity research, and has already been used several times for evaluating user-generated innovation ideas (Blohm et al. 2011a; Kristensson et al. 2004; Matthing et al. 2006; Piller and Walcher 2006). In our case, the jury consisted of 11 referees, who were either university professors in information systems, employees of SAP's marketing and R&D department, or the German SAP University Competence Centers. Idea quality was measured with four items that are internally used by SAP and reflect the dimensions of novelty, relevance, feasibility, and elaboration. For evaluation, the idea descriptions were copied into separate evaluation forms which were randomized and contained the scales for idea evaluation as well. The forms were handed out to the referees, which were assigned to rate the ideas with the four items on a rating scale from 1 (lowest) to 5 (highest) independently from the other referees. We assessed the Intra-Class-Correlation-Coefficients (ICC) of the expert evaluations that should exceed the value of 0.7 (Amabile 1996). We considered this as met for all items excluding feasibility whose ICC was 0.5. Based on the mean quality scores of the ideas, we calculated an aggregated quality ranking.

Experiment Stage I: Evaluation of Prediction Market Designs

In this section, we discuss and analyze the six market configurations according to the 2x3 factorial design of stage I in order to identify a robust market to which we compare the multi-criteria scale in stage II.

Prediction Market Configurations

The two factors investigated in stage I of the experiment are the liquidity of the market maker and the overall design of the market. Both factors are fundamentally influencing the mechanics of prediction markets, and the feedback these information systems provide to its users. As system feedback highly influences the usage of information systems (Bajaj and Nidumolu 1998; Kim and Malhotra 2005), these two factors are very likely to affect trading behavior and market accuracy as well as market perception.

The effective liquidity of the Logarithmic Market Scoring Rules (LMSR) market maker can easily be adjusted with an elasticity constant b . For designers of prediction markets choosing an appropriate value for b is a difficult issue (Berg and Proebsting 2009; Pennock and Sami 2008). Too small values of b create highly volatile markets where prices change very dynamically – even for small transactions. By contrast, high values of b create stiff markets in which prices hardly move. Whereas Berg and Proebsting (2009) offer an approach for estimating appropriate values for b , e.g., based on the number of traders and the amount of their seed capital, it is not yet known how evaluation accuracy and user satisfaction are affected by varying values of b . As participation in OI communities may change very dynamically it is pivotal to know how users react to different liquidity settings which highly determine system feedback that is provided to users. In high liquidity settings, users face more extreme situations in which high profits and high losses can be generated and thus may stimulate user action. Thus, high liquidity settings may improve rating accuracy as they motivate users to monitor their trades and the market in general more actively. However, from a viewpoint of cognitive load theory, high liquidity settings might be less preferable as such dynamic markets might be harder to interpret hampering evaluation accuracy of users. We tested three different liquidity settings simulating a low ($b=877$; assuming 80 traders), a moderate ($b=548$; 50 traders), and a high liquidity of the market maker ($b=219$; 20 traders) in relation to the traders in each market.

Contract and market design is fundamental for the success of prediction markets (Wolfers and Zitzewitz 2004). In our case, 24 ideas basically represent 24 different events that can be traded. Such non-binary event spaces have been implemented in two different ways, and is not yet clear which alternative might be more appropriate for the use context of OI communities. Most researchers set up a single market containing more than two tradable events (Gaspoz and Pigneur 2008; Soukhoroukova et al. 2011; Stathel et al. 2009), i.e., all contracts for all ideas are traded on one market that is run by one market maker. In these markets, traders are able to hold stocks in their portfolio of which they think the underlying event will occur at the market end (we call this 'single-markets'). Contrary, it is also possible to set up a single market for each tradable event or idea ('multi-markets') (Bothos et al. 2009). In these markets, each idea is represented by two contracts that we call TOP-contracts ('the idea will be one of the five best ideas among all ideas on the market') and FLOP-contracts ('the idea will *not* be one of the five best ideas among

all ideas on the market’). The single markets are unified via a common user interface so that they appear as one market to the user. Each of these markets is operated by a separate market maker. Due to the different numbers of market makers, single- and multi-markets will deliver varying system feedback from a user’s point of view. Whereas every trade on single-markets influences the prices of all other idea contracts, trades on the multi-market affect only the market price of an idea contract’s counterpart (i.e., the FLOP contract for a given TOP-contract). From a perspective of cognitive load theory, the two market designs lead to different facets of complexity inducing cognitive load whose effects are not yet researched. Whereas single-markets may tend to induce higher extraneous cognitive load by continuously updating all prices, multi-markets employ a more complex market design increasing intrinsic cognitive load.

Analysis of Prediction Market Configurations

In order to identify the best performing prediction market, we analyzed their accuracy on an aggregated level (cf. Table 1). We checked whether there is a statistically significant concurrence between the markets and the expert evaluation, calculating their Kendall-Tau rank-order correlations, and Mean Absolute Percentage Errors (MAPE) (Armstrong and Collopy 1992). We used the ranking of ideas according to their prices, and compared it to the ranking produced by the expert ratings. The multi-market design with medium liquidity of the market maker (PM5) has the highest correlation with the expert rating ($p < 0.05$), and the third lowest MAPE (smaller is better as it is a error measure) only slightly above than the smallest MAPE (9%). As described in stage II of our experiment, we applied exploratory and confirmatory factor analysis in order to validate our satisfaction scale. As the values were almost identical to those in stage II of our experiment, we do not report them here. Applying Analysis of Variance reveals that there are only little difference between the individual configurations ($F_{5,375} = 0.73$; $p = n.s.$). However, PM5 tends to be the most satisfying prediction market for users.

		PM1	PM2	PM3	PM4	PM5	Experts	MAPE	Satisfaction
Single-Market Design	High Liquidity (PM1)	-					-0.01	1.77	2.85
	Medium Liquidity (PM2)	0.29*	--				0.14	1.79	2.91
	Low Liquidity (PM3)	0.49**	0.29*	-			0.22	1.22	3.06
Multi-Market Design	High Liquidity (PM4)	0.47**	0.38*	0.44**	-		0.02	1.89	3.03
	Medium Liquidity (PM5)	0.49**	0.37**	0.37**	0.52**	-	0.33*	1.31	3.09
	Low Liquidity (PM6)	0.27	0.05*	0.21**	0.30*	0.19*	0.03	1.24	2.98

*significant with $p < 0.05$; **significant with $p < 0.01$

Conclusion

We choose PM5 for our comparison with the multi-criteria rating scale as it performs best in terms of evaluation accuracy and evaluation satisfaction on a general level. A more detailed discussion of the results of experimental phase I can be found in Appendix B.

Experiment Stage II: Prediction Market vs. Multi-Criteria Scale

In this section, we use the best performing market from experimental stage I to test our research model and compare it to a multi-criteria rating scale that performed best in previous research (Riedl et al. 2010).

Construct Validation

In order to confirm validity and reliability of our constructs, we applied exploratory and confirmatory factor analysis using SPSS and AMOS 19 (cf. Table 2). All items loaded unambiguously on the two factors that can clearly be interpreted. Multivariate normality was confirmed and therefore Maximum-Likelihood-Estimation applied. Composite Reliabilities (CR) exceeded values of 0.5, and Average Variance Explained (AVE) for each factor was at least 0.5, and thus convergent validity could be assumed (Bagozzi

and Yi 1988). The discriminant validity 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 and Larcker 1981). The factors' squared multiple correlation was 0.01, so that discriminant validity could be assumed. Cronbach Alphas of at least 0.73 suggest good reliability of factors (Nunnally and Bernstein 1994). However, we eliminated item PTV3 due to a low Individual Item Reliabilities (IIR) of 0.31 as this is far below the minimum threshold of 0.4 (Bagozzi and Yi 1988). Finally, we checked the global fit of our measurement model. The χ^2 -test was significant ($p = 0.01$), but the χ^2 / df -ratio (1.86) was well below the upper threshold of 5.00 (Wheaton et al. 1977). Global fit measures suggested excellent fit as well: GFI = 0.98 (≥ 0.9), AGFI = 0.95 (≥ 0.9), NFI = 0.91 (≥ 0.95), CFI = 0.96 (≥ 0.95), RMSEA = 0.06 (≤ 0.06) and SRMR = 0.06 (≤ 0.11) (Browne and Cudeck 1993; Bühner 2008).

Table 2. Factor Analysis of Constructs

	Item	Factor		α	IIR	CR	AVE
		Satisfaction	PTV				
SAT4	I feel happy with my idea transactions / evaluations.	0.88	-0.03	0.85	0.77	0.85	0.60
SAT3	I feel confident that my idea transactions / evaluations are correct.	0.86	0.04		0.58		
SAT1	I feel satisfied with my idea transactions / evaluations.	0.83	0.04		0.41		
SAT2	Trading / Evaluating the ideas met my expectations.	0.76	-0.09		0.64		
PTV2	To what extent did you come up against unexpected factors in trading / evaluating the ideas?	0.04	0.77	0.73	0.45	0.78	0.48
PTV1	To what extent did you come across problems about which you were unsure while trading / evaluating ideas?	0.07	0.77		0.46		
PTV4	To what extent do you feel that it is difficult to trade / evaluate the ideas?	-0.03	0.74		0.38		
PTV3	To what extent do you feel that your trades / evaluations were vague and difficult to anticipate?	-0.14	0.69		0.31		
	Eigenvalues (Variance Explained in %)	2.81 (35.1)	2.20 (27.5)				

MSA = 0.73; Bartlett-test of specificity: $\chi^2 = 354.24$, $p = 0.000$; principal component analysis; varimax-rotation; $n = 132$. The bold values indicate the attribution of the variables to one of the two factors.

Hypothesis Testing

We tested our research model using the best-performing prediction market from stage I of the experiment and a multi-criteria scale that was superior in previous research (Riedl et al. 2010). First, we performed an analysis on the aggregated level. The individual user ratings in the rating scale treatment were aggregated, and an idea ranking was constructed according to the ideas' mean quality scores. This resulted in a rank-ordered list of the ideas according to their quality evaluation, similar to the price-based ranking of the prediction market. The multi-criteria scale achieved a Kendall-Tau correlation with the expert rating of 0.41 (significant with $p < 0.01$) and a MAPE of 0.59. There was a correlation of 0.30 ($p < 0.05$) between the price ranking of the prediction market and the ranking produced by the multi-criteria scale.

In order to test our hypotheses, we need to switch the level of analysis from the aggregated level to the individual participant, where we constructed a 'fit measure' indicating how well an individual has performed in terms of evaluation accuracy. In the context of OI communities, it can be assumed that a

participant’s evaluations are accurate if he or she is able to effectively identify the ‘best’ ideas. However, this true idea quality is a priori unknown, and the community evaluations can only serve as a pre-selection for a further internal review phase (Berg-Jensen et al. 2010; Di Gangi and Wasko 2009). Thus, it is pivotal that the best ideas are identified correctly by the participants (Reinig et al. 2007). In creativity research, judgmental accuracy of laypersons is usually determined by assessing the concurrent validity of their judgments compared to those of an expert jury, e.g., by counting correctly identified ideas (Runco and Basadur 1993; Runco and Smith 1992). In order to measure a single user’s evaluation accuracy we adapted the approach of Riedl et al (2010). Accordingly, we defined the best five (ca. 21%) and eight ideas (ca. 33%) from the high quality sample strata as ‘good ideas’ and the worst five and eight ideas from the low quality sample strata as ‘bad ideas.’ We chose this cut off criteria as about 10-30% of user-generated innovation ideas are of high quality (Blohm et al. 2011a; Franke and Hienerth 2006; Walcher 2007). For each user, we counted the correctly classified ideas. On the prediction market, we considered an idea to be correctly classified, when users had TOP-contracts of the five (eight) ‘good’ ideas and FLOP-contracts of the five (eight) ‘bad’ ideas in their final portfolios. For the multi-criteria rating scale, we followed a similar approach in which we counted the best five (eight) ideas which received a rating higher than the mean rating of that idea and we counted the worst five (eight) ideas which received a rating lower than the mean. Both values were then corrected with their error by subtracting the number of misclassifications (‘good ideas’ classified as ‘bad’ and vice versa). Additionally, we normalized fit scores with the number of evaluated ideas as participants did not have to evaluate a fixed number of ideas. We performed our analysis using both cut off criteria. As they lead to almost identical results, we report only the results that are based on the more severe five idea ratio, as we believe that this reflects reality most closely.

We followed the recommendations of Cohen et al. (2003) and applied moderated, hierarchical ordinary least square (OLS) regression to test the moderating effects of PTV (H3a and H3b). As this procedure requires a test of direct effects of the evaluation mechanism used on evaluation accuracy (H1) and evaluation satisfaction (H2), we also used OLS regression to test H1 and H2. This represents an objective, indirect measurement of cognitive load (Brünken et al. 2003). Each hypotheses group regarding evaluation accuracy (H1 and H3a) and evaluation satisfaction (H2 and H3b) was tested in a single regression model. Because ‘evaluation mechanism use’ has categorical measurement level (prediction market; rating scale), we applied dummy coding in which the prediction market served as reference group (West et al. 1996). We used factor scores so that there was no need for standardization estimating the following regression equation (Frazier et al. 2004):

$$Y = b_0 + b_1 \text{ Evaluation Mechanism Use (Dummy)} + b_2 \text{ PTV} + b_3 \text{ PTV} \times \text{ Evaluation Mechanism Use (Dummy)}$$

We ran regression analyses on rating accuracy and rating satisfaction with PTV as an independent variable. PTV had no direct effects on evaluation accuracy and satisfaction. In the second step, we entered the dummy variable ‘Evaluation Mechanism Use’ to test H1 and H2. The dummy had a positive, statistically significant influence on both dependent variables (cf. Table 3). 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 term is entered (West et al. 1996). Including the dummy variable in the regression, a significant gain in explained variance can be detected for evaluation accuracy (cf. Table 3) and satisfaction (cf. Table 4). Thus, all four hypotheses can be supported. Finally, we estimated means for evaluation accuracy and satisfaction for representative groups who scored one standard deviation below and above the mean on the predictor and moderator variables in order to interpret the interaction effects (cf. Figure 4 and 5) (Cohen et al. 2003; Frazier et al. 2004).

Table 3. Moderated Regression Results on Evaluation Accuracy

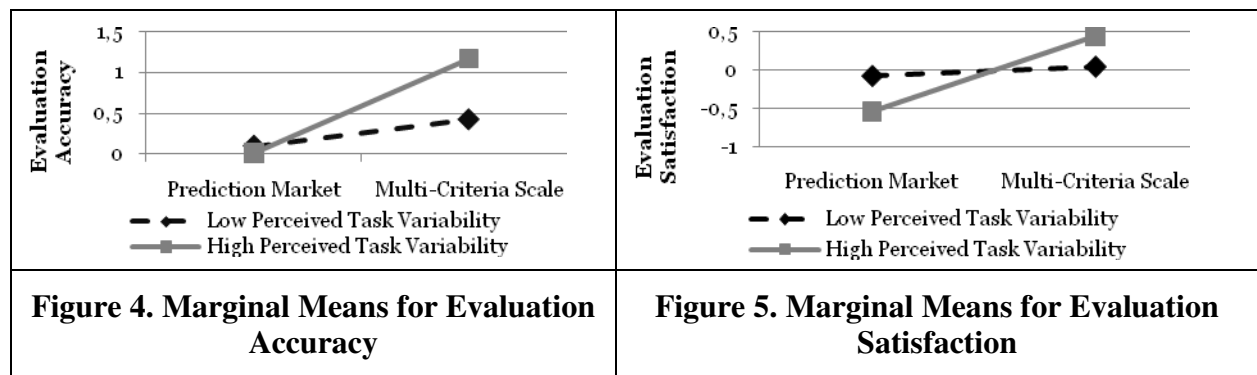
Step	Independent Variable	B	β	R ²	ΔR ²	Hypotheses	Supported
1	Perceived Task Variability	-0.04	-0.04	0.04	-		
2	Evaluation Mechanism Use (Dummy)	0.74***	0.36***	0.17	0.13***	H1	Yes
3	Perceived Task Variability x Evaluation Mechanism Use (Dummy)	0.41*	0.28*	0.21	0.04***	H3a	Yes

N = 132, *** significant with $p < 0.001$, ** significant with $p < 0.01$, * significant with $p < 0.05$

Table 4. Moderated Regression on Evaluation Satisfaction

Step	Independent Variable	B	β	R ²	ΔR^2	Hypotheses	Supported
1	Perceived Task Variability	-0.27	-0.27	0.00	-		
2	Evaluation Mechanism Use (Dummy)	0.53*	0.27*	0.08	0.08***	H2	Yes
3	Perceived Task Variability x Evaluation Mechanism Use (Dummy)	0.36**	0.26**	0.11	0.03**	H3b	Yes

N = 132, *** significant with $p < 0.001$, ** significant with $p < 0.01$, * significant with $p < 0.05$



Summary and Discussion

The first stage of our experiment compared six configurations of a prediction market (single-market and multi-market design with three different liquidity settings each). We found the multi-market design with medium liquidity (PM5) to result in the highest Kendall-tau correlation with the base-line expert rating and the third lowest MAPE. PM5's correlation of 0.33 lies also in the range reported by other researchers of 0.43 (LaComb et al. 2007) and of 0.10, 0.39 and 0.47 (Soukhoroukova et al. 2011). A possible explanation for our slightly lower correlations could be that both other studies were conducted inside company boundaries with employees and not with external novices, as in OI communities. However, PM5 can be seen as a robust, best-of breed prediction market set-up, and was thus compared to a multi-criteria scale in the second stage of our experiment. Using questionnaire data, system-captured experiment data, and an independent expert evaluation of idea quality, the proposed model was tested. It was expected that the rating scale would lead to both higher evaluation accuracy and higher evaluation satisfaction compared to that of the prediction market. Both hypotheses are supported (H1 and H2). Moreover, it was expected that PTV would have a moderating effect on the relationship between the evaluation mechanism use and evaluation accuracy, as well as the users' evaluation satisfaction. These hypotheses were supported (H3a and H3b). We also tested for a direct effect of PTV on evaluation accuracy and evaluation satisfaction, but no support was found.

While advocates of prediction markets argue that efficient markets exhibit great potential in aggregating dispersed information, our analysis suggests that while the general argument might be true for various other contexts, the relative performance of rating scales in the context of community-based idea evaluation is significantly higher ($p < 0.001$). The measurements of both the individual user's evaluation accuracy measured by the fit-score as well as the aggregated idea ranking based on both MAPE and Kendall-Tau correlation support this finding. This can be grounded in three reasons. (1) As idea quality on its own is already a fuzzy concept, using a more direct method of assessing idea quality could be beneficial. Assessing a complex concept, such as idea quality, through an indirect price building of prediction markets might simply add too much 'noise' in order to arrive at valid idea rankings. (2) The evaluation mechanism of a market might be too complex, and trading ideas might distract users from developing mental representations of the actual idea quality that are the foundations of evaluating idea

quality accurately. In this regard, the working memory of prediction market users could have been already fully loaded handling the prediction market so that existing knowledge that is necessary for idea evaluation could not be accessed. These negative effects might offset potential benefits of the prediction market mechanism in integrating dispersed information leading to lower evaluation accuracy. (3) Properly designed multi-criteria scales may stimulate decision making of raters so that these can develop a fuller understanding of the problem integrate aspects into their judgment they would not have thought of otherwise (Keeney 1992). All three reasons lead to the conclusion that the use of the considerably more complex evaluation mechanism of a prediction market is not warranted by superior accuracy. Users of prediction markets are at higher risk of facing a situation of cognitive overload which hampers the ability of evaluating ideas accurately than participants that assess idea quality with rating scales. Further, the moderating effect of PTV gives strong support to our assumptions. Generally, users perceiving the experimental task as highly variable are more endangered by the risk of cognitive overload, as the given task seems more challenging to them. Accordingly, these users are affected in a stronger way by a declining ability of evaluating ideas correctly on prediction markets than are users facing low PTV.

As user satisfaction is of great importance in online communities that rely on voluntary participation and contribution, evaluation accuracy of a given evaluation mechanism is not the only selection criteria for companies operating OI communities. The multi-criteria scale induced a significantly higher degree of user satisfaction than the best-performing prediction market from stage I. In this context, the higher mechanism complexity of prediction markets may have induced higher decisional stress so that more users were facing a state of-hyper-vigilance, leading to lower decision satisfaction. Our results also show that this effect is stronger for users perceiving a high PTV and high cognitive load.

In summary, a combination of a web-based experiment, questionnaire data, and expert rating provides insights that would not have been possible with only one source of data, and thus offers a more detailed understanding of evaluation mechanisms for OI communities. Overall, there is mutual support between the different analysis methods and data sources to suggest that rating scales lead to both higher evaluation accuracy and evaluation satisfaction. This effect is strengthened even more when taking PTV into consideration.

Conclusion

Theoretical Contribution

From a theoretical perspective, this paper is a first attempt to integrate two separate streams of research. While rating scales (e.g., Berg-Jensen et al. 2010; Riedl et al. 2010) and prediction markets (e.g., LaComb et al. 2007; Soukhoroukova et al. 2011) have been studied as means to aggregate opinions of crowds, this research is the first to compare the relative performance of both mechanisms as means for idea evaluation in OI communities – not only on the aggregated mechanism level but also on the level of the individual user. Based on the analysis of the user level, we were able to explain performance differences between the two investigated mechanisms. We applied cognitive load theory in a new research context. Our research contributes to theory building in the emerging area of collective intelligence by extending our understanding of how and why such mechanisms do work (Zwass 2010). It contributes to the field of market engineering research, where it enables a better understanding of how prediction market designs affect trading outcomes and may generalize to various related and yet not answered research questions. For instance, it may help to explain why prediction markets are only sparsely used, as people do not understand how these markets work (Graefe 2009; Kamp and Koen 2009). Our analysis of PTV offers a contribution to our understanding of task complexity as key elements influencing IT system usage by groups, in which collective judgment tasks have rarely been studied (Zigurs and Buckland 1998). Finally, our research contributes by suggesting novel integration mechanisms for external knowledge that can be used to improve organizations' absorptive capacity by engaging a larger user base (Lewin et al. 2011).

Methodological Contribution

This paper reports results of a multi-stage experiment combining multiple research methods (web-experiment, questionnaire, expert evaluation), and two levels of analysis (mechanism and user level). This follows recommendations to use method triangulation to avoid common method bias (Sharma et al. 2009) and a call for more advanced experimental designs (Shadish et al. 2002). Our multi-

treatment/multi-method approach underlines the robustness of our results. Our extension of the fit-score method of Riedl et al. (2010) offers a means to dissect compound results which are initially only available on the aggregated level. Using this fit-score, aggregated observation can be dissected and brought down to the user level where it can be used to perform a richer analysis by combining the system-captured behavior data with questionnaire data about perceived usage that may directly influence user behavior. Thus, the fit-score method helps to entangle complex research results, and enhances our ability to give detailed design recommendations regarding the construction of IT artifacts (Benbasat and Zmud 2003).

Practical Contribution

From a practical perspective, recommendations regarding the design of mechanisms for community-based judgments of idea quality can be given. The effective and accurate design of mechanisms for collective judgment tasks is of paramount importance to help organization to overcome their limited absorptive capacity by outsourcing idea evaluation to a crowd of users. Based on our results, we argue that a multi-criteria scale is superior for community-based idea evaluation in terms of evaluation accuracy and satisfaction. Regarding prediction markets using the LMSR market maker, there were no significant differences regarding evaluation satisfaction on the tested market designs. Evaluation accuracy can be maximized with a medium liquidity of the market maker on the 'multi-market' design. On a general level, the strong moderating effect of PTV suggests that users will strongly react to information and functionalities provided during the process of idea evaluation. Thus, designers of community-based idea evaluation mechanisms should use these very carefully in order to avoid overloading participants cognitively. Consequently, our design recommendation offers valuable suggestions for how a community-based idea quality assessment through a multi-criteria rating scale can be used to supplement or even replace expert panels in their assessment of customer-generated ideas.

Research Limitations and Potential Future Research Directions


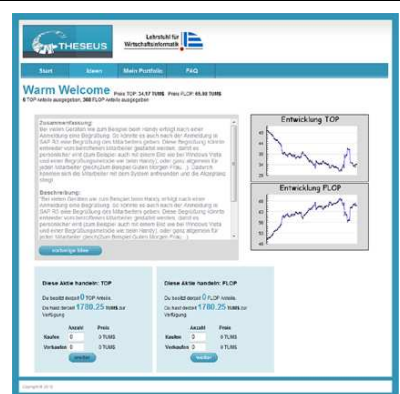
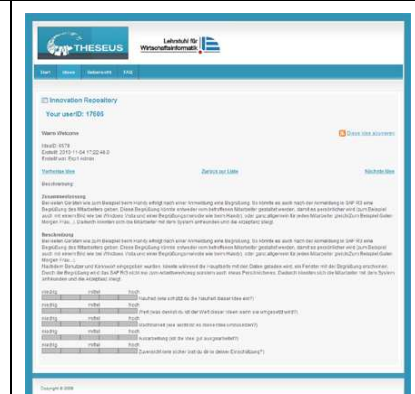
Some general shortcomings resulting from conducting a controlled experiment apply to our research. While our web-based experiment was intended to closely reflect actual community behavior, general threats to the external validity have to be acknowledged resulting from the use of students. Furthermore, following the 'wisdom of the crowd' paradigm, the expert rating which served as the baseline for all our relative comparisons might be deficient, although experts generally outperform non-experts (see Ericsson and Lehmann 1996 for a review). A small group of experts might be more prone to a fixed mind-set rather than a broader community, and thus certain aspects of some ideas might have been overlooked. However, as true idea quality is not directly observable, assessment of idea quality through an expert panel is generally performed for idea selection. Even when accepting expert judgment as biased, our results retain their validity insofar as they provide insights into how an expert panel can be supported or maybe even replaced by a community in order to conserve valuable resources and to cope with increasing numbers of submissions which can no longer be manually assessed by a small group of experts. As the ideas in our experiment derived from the domain of software development our findings should be replicated with ideas from other contexts to ensure generalizability. However, we think that our results are generally applicable for textual idea descriptions. There are OI communities in which ideas are not based on textual descriptions but rather resemble visual designs (Berg-Jensen et al. 2010; Bullinger et al. 2010). Dual coding theory (Paivio 1986) suggests that verbal and visual cues are processed differently in human cognition so that differences in idea evaluation are likely to occur. Moreover, prediction markets for idea evaluation generally suffer from the fact that no real observable outcome exists, to which payoffs can be tied. Participants could bet on expert evaluations and not on idea quality itself

Our research found that users perceive the investigated evaluation mechanisms cognitively different and this perception highly influences the mechanism's outcome. Thus, a more indulgent understanding of user cognitions is necessary to design more powerful mechanisms for group judgments and social interaction systems in general. In this regard, future research should especially consider the decision process of idea evaluators. Understanding this process, mechanisms can be tailored to deliver higher decision support and better evaluation accuracy and satisfaction.

Acknowledgement

This research received funding through the GENIE project by the German Ministry of Research and Education (BMBF) under contract No. FKZ 01FM07027 and the German Ministry of Economics and Technology (BMWi) under grant code 01MQ07024. The second author also acknowledges supported by the German Research Foundation under grant code RI 2185/1-1. The responsibility for the content of this publication lies with the authors.

Appendix A – Idea Evaluation Mechanisms

		
<p>Transaction form on the details page of an idea: users can enter the quantity of idea contracts to be bought (sold) in the transaction form to calculate the buying (selling) price; the graph depicts price trends of the idea contract</p>	<p>Transaction form on the details page of an idea: transaction form resembles single market design, however there is a separate form for TOP and FLOP-contracts, the graph depicts price trends of TOP and FLOP contract</p>	<p>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”.</p>
<p>‘Single-Market’ Prediction Market</p>	<p>‘Multi-Market’ Prediction Market</p>	<p>Multi-Criteria Rating Scale</p>

Appendix B – Extended Discussion of Experimental Stage I

The results of experimental phase I show that prediction markets with a high liquidity setting of the market maker result in lowest correlations and highest MAPEs irrespective of the market design whereas medium liquidity worked best on the multi-market and a low liquidity setting on the single-market design. These results suggest that prediction markets with medium to low liquidity settings tend to be more accurate than low liquidity settings. These findings can easily be interpreted in terms of cognitive load theory and PTV. The higher the liquidity setting of the market maker the more contract prices change on the market. This increase in market volatility may make interpretation of the financial data that is necessary for trading successfully more complicated resulting in higher cognitive load and lower rating accuracy.

In regard to the market design the multi-market designs tends to produce more accurate results than the single-market design. A reason for this may be a higher decision support as users can buy idea contracts of which they think they represent the best and the worst ideas instead of idea contracts that reflect the best ideas only. It can be argued that more complex design of the multi-market may induce a lower cognitive load as it better fits the process of idea evaluation during which filtering out the worst ideas is a fundamental process step. Additionally, the multi-market design creates more stable prices of idea contracts making interpretation of results easier.

References

- Amabile, T.M. 1996. *Creativity in Context. Update to Social Psychology of Creativity*, (1 ed.). Oxford: Westview Press.
- Armstrong, J.S., and Collopy, F. 1992. "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons" *International Journal of Forecasting* (8:1), pp. 69-80.
- Arrow, K.J., Forsythe, R., Gorham, M., Hahn, R., Ledyard, J.O., Levmore, S., Litan, R., Milgrom, P., Nelson, F.D., Neumann, G.R., Ottaviani, M., Schelling, T.C., Shiller, R.J., Smith, V.L., Snowberg, E., Sunstein, C.R., Tetlock, P.C., Tetlock, P.E., Varian, H.R., Wolfers, J., and Zitzewitz, E. 2008. "The Promise of Prediction Markets," *Science* (320:5878), pp. 877-878.
- Baddeley, A. 1986. *Working Memory*, (1 ed.). Oxford, UK: Oxford University Press.
- Bagozzi, R.P., and Yi, Y. 1988. "On the Evaluation of Structural Equation Models," *Journal of the Academy of Marketing Sciences* (16:1), pp. 74-94.
- Bajaj, A., and Nidumolu, S.R. 1998. "A Feedback Model to Understand Information System Usage," *Information & Management* (33:4), pp. 213-224.
- Benbasat, I., and Zmud, R. 2003. "The Identity Crisis within the IS Discipline: Defining and Communicating the Discipline's Core Properties," *MIS Quarterly* (27:2), pp. 183-194.
- Berg-Jensen, M., Hienert, C., and Lettl, C. 2010. "Forecasting the Attractiveness of User-Generating Designs Via Online-Data," in: *2010 Academy of Management Annual Meeting*. Montreal, Canada.
- Berg, H., and Proebsting, T.A. 2009. "Hanson's Automated Market Maker," *The Journal of Prediction Markets* (3:1), pp. 45-59.
- Berg, J.E., and Rietz, T.A. 2003. "Prediction Markets as Decision Support Systems," *Information Systems Frontiers* (5:1), pp. 79-93.
- Berthold, A., and Jameson, A. 1999. "Interpreting Symptoms of Cognitive Load in Speech Input," in: *Proceedings of the seventh international conference on User modeling*. Banff, Canada: Springer-Verlag New York, Inc.
- Blohm, I., Bretschneider, U., Leimeister, J.M., and Krcmar, H. 2011a. "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), pp. 106-122.
- Blohm, I., Köroglu, O., Leimeister, J.M., and Krcmar, H. 2011b. "Absorptive Capacity for Open Innovation Communities - Learnings from Theory and Practice," in: *2011 Academy of Management Annual Meeting*. San Antonio, Texas / USA.
- Blume, M., Luckner, S., and Weinhardt, C. 2010. "Fraud Detection in Play-Money Prediction Markets," *Information Systems and E-Business Management* (8:4), pp. 395-413.
- Boer, K., Kaymak, U., and Spiering, J. 2007. "From Discrete-Time Models to Continuous-Time Asynchronous Modeling of Financial Markets," *Computational Intelligence* (23:2), pp. 142-161.
- Bogers, M., Afuah, A., and Bastian, B. 2010. "Users as Innovators: A Review, Critique, and Future Research Directions," *Journal of Management* (36:4), pp. 857-875.
- Bonabeau, E. 2009. "Decision 2.0: The Power of Collective Intelligence," *MIT Sloan Management Review* (50:2), pp. 44-52.
- Bothos, E., Apostolou, D., and Mentzas, G. 2009. "Collective Intelligence for Idea Management with Internet-Based Information Aggregation Markets," *Internet Research* (19:1), pp. 26-41.
- Botti, S., and Iyengar, S.S. 2006. "The Dark Side of Choice: When Choice Impairs Social Welfare," *Journal of Public Policy & Marketing* (25:1), pp. 24-38.
- Boudreau, M.-C., Gefen, D., and Straub, D.W. 2001. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1), pp. 1-16.
- Bretschneider, U. 2011. "Die Ideen Community Zur Integration Von Kunden in Die Frühen Phasen Des Innovationsprozesses. Empirische Analysen Und Implikationen Für Forschung Und Praxis," in: *Chair of Information Systems (I17)*. Garching b. München: Technische Universität München.
- Browne, M.W., and Cudeck, R. 1993. "Alternative Ways of Assessing Model Fit," in *Testing Structural Equation Models*, K.A. Bollen and J.S. Long (eds.). Newbury Park, CA, USA: Sage.
- Brünken, R., Plass, J.L., and Leutner, D. 2003. "Direct Measurement of Cognitive Load in Multimedia Learning," *Educational Psychologist* (28:1), pp. 53-61.
- Bühner, M. 2008. *Einführung in Die Test- Und Fragebogenkonstruktion*, (2 ed.). München, Germany: Pearson Studium.

- Bullinger, A., Neyer, A.K., Rass, M., and Möslein, K. 2010. "Community-Based Innovation Contests: Where Competition Meets Cooperation," *Creativity & Innovation Management* (19:3), pp. 290-303.
- Cady, S.H., and Valentine, J. 1999. "Team Innovation and Perceptions of Consideration. What Difference Does Diversity Make?," *Small Group Research* (30:6), pp. 730-750.
- Caroff, X., and Besançon, M. 2008. "Variability of Creativity Judgments," *Learning and Individual Differences* (18:4), pp. 367-371.
- Chalmers, P.A. 2003. "The Role of Cognitive Theory in Human-Computer Interface," *Computers in Human Behavior* (19:5), pp. 593-607.
- Chen, L., Goes, P., Marsden, J.R., and Zhang, Z. 2009-10. "Design and Use of Preference Markets for Evaluation of Early Stage Technologies," *Journal of Management Information Systems* (26:3), pp. 45-70.
- Chen, Y., Chu, C.-H., Mullen, T., and Pennock, D.M. 2005. "Information Markets Vs. Opinion Pools: An Empirical Comparison," in: *6th ACM Conference on Electronic Commerce*. Vancouver, BC, Canada: ACM.
- Chen, Y., Dimitrov, S., Sami, R., Reeves, D.M., Pennock, D.M., Hanson, R.D., Fortnow, L., and Gonen, R. 2010. "Gaming Prediction Markets: Equilibrium Strategies with a Market Maker," *Algorithmica* (58:4), pp. 930-969.
- Chesbrough, H.W. 2006. *Open Innovation. The New Imperative for Creating and Profiting from Technology*, (1 ed.). Boston MA:
- Christian, L.M., Dillman, D.A., and Smyth, J.D. 2007. "Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys," *Public Opinion Quarterly* (71:1), pp. 113-125.
- Christiansen, J.D. 2007. "Prediction Markets: Practical, Experiments in Small Markets and Behaviors Observed," *The Journal of Prediction Markets* (1:1), pp. 17-41.
- Cohen, J., Cohen, P., West, S.G., and Aiken, L.S. 2003. *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, (3 ed.). Hillsdale, NJ, USA: Erlbaum.
- Cohen, W., M., and Levinthal, D., A. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation " *Administrative Science Quarterly* (35:1), pp. 128-152.
- Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., and Riedl, J. 2003. "Is Seeing Believing?: How Recommender System Interfaces Affect Users' Opinions," *SIGCHI conference on Human factors in computing systems*, Ft. Lauderdale, Florida, USA: ACM.
- Couper, M.P., Conrad, F.G., and Tourangeau, R. 2007. "Visual Context Effects in Web Surveys," *Public Opinion Quarterly* (71:4), pp. 623-634.
- Cowan, N. 2005. *Working Memory Capacity, Essays in Cognitive Psychology*. New York, NY, USA: Psychology Press.
- Cyr, D., Head, M., Larios, H., and Pan, B. 2009. "Exploring Human Images in Website Designs: A Multi-Method Approach," *MIS Quarterly* (33:3), pp. 530-566.
- Daft, R.L., and Macintosh, N.B. 1981. "A Tentative Exploration into the Amount and Equivocality of Information Processing in Organizational Work Units," *Administrative Science Quarterly* (26:2), pp. 207-224.
- Dahan, E., Soukhoroukova, A., and Spann, M. 2010. "New Product Development 2.0: Preference Markets How Scalable Securities Markets Identify Winning Product Concepts & Attributes," *Journal of Product Innovation Management* (27:2), pp. 937-954.
- Das, S. 2005. "A Learning Market-Maker in the Glosten-Milgrom Model," *Quantitative Finance* (5:2), pp. 169-180.
- Dean, D.L., Hender, J.M., Rodgers, T.L., and Santanen, E.L. 2006. "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," *Journal of the Association for Information Systems* (7:10), pp. 646-698.
- DeStefano, D., and LeFevre, J.-A. 2007. "Cognitive Load in Hypertext Reading: A Review," *Computers in Human Behavior* (23), pp. 1616-1641.
- Di Gangi, P.M., and Wasko, M.M. 2009. "Steal My Idea! Organizational Adoption of User Innovations from a User Innovation Community: A Case Study of Dell Ideastorm," *Decision Support Systems* (48:1), pp. 303-312.
- Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, (1 ed.). Cambridge, Mass., USA: Cambridge University Press.

- Ebner, W., Leimeister, J.M., and Krcmar, H. 2009. "Community Engineering for Innovations: The Ideas Competition as a Method to Nurture a Virtual Community for Innovations," *R&D Management* (39:4), pp. 342-356.
- Enkel, E., Perez-Freije, J., and Gassmann, O. 2005. "Minimizing Market Risks through Customer Integration in New Product Development: Learning from Bad Practice," *Creativity and Innovation Management* (14:4), pp. 425-437.
- Eppler, M.J., and Mengis, J. 2004. "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, Mis, and Related Disciplines," *The Information Society: An International Journal* (20:5), pp. 325-344.
- Ericsson, K.A., and Lehmann, A.C. 1996. "Experts an Exceptional Performance: Evidence of Maximal Adaption to Task Constraints," *Annual Review of Psychology* (47:1), pp. 273-305.
- Fama, E.F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance* (25:2), pp. 383-417.
- Fama, E.F. 1991. "Efficient Capital Markets: Ii," *The Journal of Finance* (46:5), pp. 1575-1617.
- Farhoomand, A.F., and Drury, D.H. 2002. "Managerial Information Overload," *Communications of the ACM* (45:10), pp. 127-131.
- Feroli, M., Dekoninck, E., Culley, S., Roussel, B., and Renaud, J. 2010. "Understanding the Rapid Evaluation of Innovative Ideas in the Early Stages of Design," *International Journal of Product Development* (12:1), pp. 67-83
- Fishbein, M. 1966. "The Relationships between Beliefs, Attitudes, and Behavior," in *Cognitive Consistency: Motivational Antecedents and Behaviorial Consequents* S. Feldmann (ed.). New York: Academic Press.
- Fornell, C., and Larcker, D.F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18:2), pp. 39-50.
- Forsythe, R., Rietz, T.A., and Ross, T.W. 1999. "Wishes, Expectations and Actions: A Survey on Price Formation in Election Stock Markets," *Journal of Economic Behavior & Organization* (39:1), pp. 83-110.
- Franke, N., and Hienert, C. 2006. "Prädikatoren Der Qualität Von Geschäftsideen: Eine Empirische Analyse Eines Online-Ideen-Forums," *Zeitschrift für Betriebswirtschaft Special Issue* (6:4), pp. 47-68.
- Franke, N., and Shah, S. 2003. "How Communities Support Innovative Activities: An Exploration of Assistance and Sharing among End-Users," *Research Policy* (32:1), pp. 157-178.
- Frazier, P.A., Tix, A.P., and Barron, K.E. 2004. "Testing Moderator and Mediator Effects in Counseling Psychology Research," *Journal of Counseling Psychology* (51), pp. 115-134.
- Gaspoz, C., and Pigneur, Y. 2008. "Preparing a Negotiated R&D Portfolio with a Prediction Market," *41st Hawaii International Conference on System Science (HICCS 41)*, Waikoloa, Big Island, Hawaii, USA.
- Gassmann, O. 2006. "Opening up the Innovation Process: Towards an Agenda," *R&D Management* (36:3), pp. 223-228.
- Goleman, D. 1996. *Emotional Intelligence. Why It Can Matter More Than Iq*, (1 ed.). London: Bloomsbury.
- Graefe, A. 2009. "Prediction Markets Versus Alternative Methods. Empirical Tests of Accuracy and Acceptability," in: *Fakultät für Wirtschaftswissenschaften*. Karlsruhe: Universität Karlsruhe (TH).
- Grise, M.-L., and Gallupe, R.B. 2000. "Information Overload: Addressing the Productivity Paradox in Face-to-Face Electronic Meetings," *Journal of Management Information Systems* (16:3), pp. 157-185.
- Gwizdka, J. 2010. "Distribution of Cognitive Load in Web Search," *J. Am. Soc. Inf. Sci. Technol.* (61:11), pp. 2167-2187.
- Haerem, T., and Rau, D. 2007. "The Influence of Degree of Expertise and Objective Task Complexity on Perceived Task Complexity and Performance," *Journal of Applied Psychology* (92:5), pp. 1320-1331.
- Hanson, R. 2003. "Combinatorial Information Market Design," *Information Systems Frontiers* (5:1), pp. 107-119.
- Hanson, R. 2007. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation," *Journal of Prediction Markets* (1:1), pp. 3-15.
- Hayek, F.A. 1945. "The Use of Knowledge in Society," *American Economic Review* (35:4), pp. 519-530.

- Healy, P.J., Linardi, S., Lowery, J.R., and Ledyard, J.O. 2010. "Prediction Markets: Alternative Mechanisms for Complex Environments with Few Traders," *Management Science* (56:11), pp. 1977-1996.
- Hwang, M.I., and Lin, J.W. 1999. "Information Dimension, Information Overload and Decision Quality," *Journal of Information Science* (25:3), June 1, 1999, pp. 213-218.
- Janis, I.L., and Mann, L. 1977. *Decision Making. A Psychological Analysis of Conflict, Choice, and Commitment*, (1 ed.). New York: The Free Press.
- Janis, I.L., and Mann, L. 1982. "A Theoretic Framework for Decision Counseling," in *Counseling on Personal Decisions: Theory and Helping on Short-Term Helping Relationships.*, I.L. Janis (ed.). New Heaven: Yale University Press.
- Jeppesen, L.B., and Frederiksen, L. 2006. "Why Do Users Contribute to Firm-Hosted User Communities? The Case of Computer-Controlled Music Instruments," *Organization Science* (17:1), pp. 45-63.
- Jian, L., and Sami, R. 2010. "Aggregation and Manipulation in Prediction Markets: Effects of Trading Mechanism and Information Distribution," in: *11th ACM conference on Electronic commerce*. Cambridge, Massachusetts, USA: ACM.
- Jokisch, M. 2007. *Active Integration of Users into the Innovation Process of a Manufacturer. The Bmw Customer Innovation Lab*, (1 ed.). München:
- Kamp, G., and Koen, P.A. 2009. "Improving the Idea Screening Process within Organizations Using Prediction Markets: A Theoretical Perspective," *The Journal of Prediction Markets* (3:2), pp. 39-64.
- Karimi, J., Somers, T.M., and Gupta, Y.P. 2004. "Impact of Environmental Uncertainty and Task Characteristics on User Satisfaction with Data," *Information Systems Research* (15:2), pp. 175-193.
- Keeney, R.L. 1992. *Value-Focused Thinking: A Path to Creative Decision-Making*, (1 ed.). Cambridge, Mass., USA: Harvards University Press.
- Kim, S.S., and Malhotra, N.K. 2005. "A Longitudinal Model of Continued Is Use: An Integrative View of Four Mechanisms Underlying Postadoption Phenomena," *Management Science* (51:5), pp. 741-755.
- Kirsh, D. 2000. "A Few Thoughts on Cognitive Overload," *Intellectica* (30:1), pp. 19-51.
- Knapp, H., and Kirk, S.A. 2003. "Using Pencil and Paper, Internet and Touch-Tone Phones for Self-Administered Surveys: Does Methodology Matter?," *Computers in Human Behavior* (19:1), pp. 117-134.
- Kristensson, P., Gustafsson, A., and Archer, T. 2004. "Harnessing the Creative Potential among Users," *The Journal of Product Innovation Management* (21:1), pp. 4-14.
- LaComb, A.C., Barnett, A., and Qimei, P. 2007. "The Imagination Market," *Information Systems Frontiers* (9:2-3), pp. 245-256.
- Lane, P.J., Koka, B.R., and Pathak, S. 2006. "The Reification of Absorptive Capacity: A Critical Review and Rejuvenation of the Construct," *Academy of Management Review* (31:4), pp. 863-883.
- LeDoux, J. 1998. *The Emotional Brain. The Mysterious Underpinning of Emotional Life*, (1 ed.). London: Weidenfeld & Nicolson.
- Leimeister, J.M. 2010. "Collective Intelligence," *Business & Information Systems Engineering* (52:4), pp. 239-242.
- Leung, R., MacLean, K., Bertelsen, M.B., and Saubhasik, M. 2007. "Evaluation of Haptically Augmented Touchscreen Gui Elements under Cognitive Load," in: *Proceedings of the 9th international conference on Multimodal interfaces*. Nagoya, Aichi, Japan: ACM.
- Lewin, A.Y., Massini, S., and Peeters, C. 2011. "Microfoundations of Internal and External Absorptive Capacity Routines," *Organization Science* (22:1), pp. 81-98.
- Limayem, M., and DeSanctis, G. 2000. "Providing Decisional Guidance for Multicriteria Decision Making in Groups," *Information Systems Research* (11:4), pp. 386-401.
- Luckner, S., and Weinhardt, C. 2007. "How to Pay Traders in Information Markets: Results from a Field Experiment," *Journal of Prediction Markets* (1:2), pp. 147-156.
- Malhotra, N.K. 1982. "Reflections on the Information Overload Paradigm in Consumer Decision Making," *The Journal of Consumer Research* (10:4), pp. 436-440.
- Malhotra, N.K. 2007. *Marketing Research. An Applied Orientation*, (5 ed.). Upper Saddle River: Pearson.
- March, J.G. 1978. "Bounded Rationality, Ambiguity and the Engineering of Choice," *The Bell Journal of Economics* (9:2), pp. 587-608.

- Matthing, J., Kristensson, P., Gustafsson, A., and Parasuraman, A. 2006. "Developping Successful Technology-Based Services: The Issue of Identifying and Involving Innovative Users," *Journal of Services Marketing* (20:5), pp. 288-297.
- Mayer, R.E., and Moreno, R. 2003. "Nine Ways to Reduce Cognitive Load in Multimedia Learning," *Educational Psychologist* (38:1), pp. 43-52.
- Miller, G.A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review* (63:2), pp. 81-97.
- Mizuyama, H., and Komatsu, T. 2010. "A Prediction Market Approach to Facilitating Consensus Building Among Supply Chain Partners," *E-Journal of Advanced Maintenance* (2:1), pp. 149-159.
- Nielsen, J. 1994. "Enhancing the Explanatory Power of Usability Heuristics," *SIGCHI Conference on Human Factors in Computing Systems*, Boston, Mass., USA, pp. 152-158.
- Nunnally, J.C., and Bernstein, I.H. 1994. *Psychometric Theory*. New York, USA: McGraw-Hill.
- O'Reilly, C.R., Parlette, G.N., and Bloom, J.R. 1980. "Perceptual Measures of Task Characteristics: The Biasing Effects of Differing Frames of Reference and Job Attitudes," *Academy of Management Journal* (23:1), pp. 118-131.
- Ogawa, S., and Piller, F. 2006. "Reducing the Risks of New Product Development," *MIT Sloan Management Review* (47:2), pp. 65-71.
- Othman, A., and Sandholm, T. 2010. "Automated Market-Making in the Large: The Gates Hillman Prediction Market," in: *11th ACM conference on Electronic commerce*. Cambridge, Mass., USA: ACM.
- Ozer, M. 2005. "What Do We Know About New Product Idea Selection?."
- Paas, F., Renkl, A., and Sweller, J. 2003. "Cognitive Load Theory and Instructional Design: Recent Developments," *Educational Psychologist* (38:1), pp. 1-4.
- Paivio, A. 1986. *Mental Representations: A Dual Coding Approach*, (1 ed.). Oxford, England: Oxford University Press.
- Pennock, D.M., and Sami, R. 2008. "Computational Aspects of Prediction Markets," in *Algorithmic Game Theory*. Cambridge: Cambridge Univ. Press, pp. 651-678.
- Perrow, C. 1967. "A Framework for the Comparative Analysis of Organizations," *American Sociological Review* (32:2), pp. 194-208.
- Piller, F.T., and Walcher, D. 2006. "Toolkits for Idea Competitions: A Novel Method to Integrate Users in New Product Development," *R&D Management* (36:3), pp. 307-318.
- Poston, R.S., and Speier, C. 2005. "Effective Use of Knowledge Management Systems: A Process Model of Content Ratings and Credibility Indicators," *MIS Quarterly* (29:2), pp. 221-244
- Reinig, B.A., Briggs, R.O., and Nunamaker Jr, J.F. 2007. "On the Measurement of Ideation Quality," *Journal of Management Information Systems* (23:4), pp. 143-161.
- Riedl, C., Blohm, I., Leimeister, J.M., and Krcmar, H. 2010. "Rating Scales for Collective Intelligence in Innovation Communities: Why Quick and Easy Decision Making Does Not Get It Right," *2010 International Conference on Information Systems*, St. Louis, Mi, USA: AIS.
- Riedl, C., May, N., Finzen, J., Stathel, S., Kaufman, V., and Krcmar, H. 2009. "An Idea Ontology for Innovation Management," *International Journal on Semantic Web and Information Systems* (5:4), pp. 1-18.
- Rochford, L. 1991. "Generating and Screening New Product Ideas," *Industrial Marketing Management* (20:4), pp. 287-296.
- Runco, M.A., and Basadur, M. 1993. "Assessing Ideational and Evaluative Skills and Creative Styles and Attitudes," *Creativity and Innovation Management* (2:3), pp. 166-173.
- Runco, M.A., and Smith, W.R. 1992. "Interpersonal and Intrapersonal Evaluations of Creative Ideas," *Personality and Individual Differences* (13:3), pp. 295-302.
- Sainfort, F., and Booske, B. 2000. "Measuring Post-Decisions Satisfaction," *Medical Decision Making* (20:1), pp. 51-61.
- Schmutz, P., Heinz, S., Métrailler, Y., and Opwis, K. 2009. "Cognitive Load in Ecommerce Applications - Measurement and Effects on User Satisfaction," *Advances in Human-Computer Interaction*.
- Schulz, C., and Wagner, S. 2008. "Outlaw Community Innovations," *International Journal of Innovation Management* (12:3), pp. 399-418.
- Servan-Schreiber, E., Wolfers, J., Pennock, D.M., and Galebach, B. 2004. "Prediction Markets: Does Money Matter?," *Electronic Markets* (14:3), pp. 243-251.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, (1 ed.). Boston, Mass., USA: Houghton Mifflin Company.

- Shannon, C.E. 1948. "A Mathematical Theory of Communication," *The Bell System Technical Journal* (27), pp. 379-423 & 623-656.
- Sharma, R., Yetton, P., and Crawford, J. 2009. "Estimating the Effect of Common Method Variance: The Method-Method Pair Technique with an Illustration from Tam Research," *MIS Quarterly* (33:3), pp. 473-490.
- Shneiderman, B., and Plaisant, C. 2004. *Designing the User Interface*. Reading, Mass., USA: Addison-Wesley.
- Slamka, C., Jank, W., and Skiera, B. 2011. "Second-Generation Prediction Markets for Information Aggregation: A Comparison of Payoff Mechanisms," *Journal of Forecasting* (forthcoming).
- Soukhoroukova, A., Spann, M., and Skiera, B. 2011. "Sourcing, Filtering, and Evaluating New Product Ideas: An Empirical Exploration of the Performance of Idea Markets," *Journal of Product Innovation Management* (forthcoming).
- Spann, M., and Skiera, B. 2003. "Internet-Based Virtual Stock Markets for Business Forecasting," *Management Science* (49:10), pp. 1310-1326.
- Speier, C., Vessey, I., and Valacich, J.S. 2003. "The Effects of Interruptions, Task Complexity, and Information Presentation on Computer-Supported Decision-Making Performance," *Decision Sciences* (34:4), pp. 771-797.
- Stathel, S., van Dinther, C., and Schönfeld, A. 2009. "Service Innovation with Information Markets," *Wirtschaftsinformatik*, Wien.
- Stewart, T., and Travis, D. 2003. "Guidelines, Standards, and Style Guides," in *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears (eds.). London: Lawrence Erlbaum Associates.
- Sweller, J. 1988. "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science* (12:2), pp. 257-285.
- Torodova, G., and Durisin, B. 2007. "Absorptive Capacity: Valuing a Reconceptualization," *Academy of Management Review* (32:3), pp. 774-786.
- Toubia, O., and Flores, L. 2007. "Adaptive Idea Screening Using Consumers," *Marketing Science* (26:3), pp. 342-360.
- Triantaphyllou, E. 2000. *Multi-Criteria Decision Making Methods: A Comparative Study*, (1 ed.). Dordrecht, The Netherlands: Kluwer.
- Urban, G.L., and Hauser, J.R. 1993. *Design and Marketing of New Products*, (1 ed.). Englewood Cliffs, NJ/USA: Prentice Hall.
- van Merriënboer, J.J.G., Kirschner, P.A., and Kester, L. 2003. "Taking the Load Off a Learner's Mind: Instructional Design for Complex Learning," *Educational Psychologist* (38:1), pp. 5-13.
- van Schaik, P., and Ling, J. 2007. "Design Parameters of Rating Scales for Web Sites," *ACM Transactions on Computer-Human Interaction* (14:1), p. 35.
- Voich, D. 1995. *Comparative Empirical Analysis of Cultural Values and Perceptions of Political Economy Issues*, (1 ed.). Westport, Connecticut: Praeger.
- von Hippel, E. 2005. *Democratizing Innovation*, (1 ed.). Cambridge, Mass.: MIT Press.
- Walcher, P.-D. 2007. *Der Ideenwettbewerb Als Methode Der Aktiven Kundenintegration*, (1 ed.). Wiesbaden: Gabler.
- West, J., and Lakhani, K. 2008. "Getting Clear About Communities in Open Innovation," *Industry and Innovation* (15:2), pp. 223-231.
- West, S.G., Aiken, L.S., and Krull, J.L. 1996. "Experimental Personality Designs: Analyzing Categorical by Continuous Variable Interactions," *Journal of Personality* (64:1), pp. 1-49.
- Wheaton, B., Muthén, B., Alwin, D.F., and Summers, G.F. 1977. "Assessing Reliability and Stability in Panel Models," in *Sociological Methodology*, D.R. Heise (ed.). San Francisco, CA, USA: Jossey-Bass.
- Winkelmann, A., Herwig, S., Pöppelbuß, J., Tiebe, D., and Becker, J. 2009. "Discussion of Functional Design Options for Online Rating Systems: A State-of-the-Art Analysis," *17th European Conference on Information Systems (ECIS 2009)*, Verona.
- Withey, M., Daft, R.L., and Cooper, W.H. 1983. "Measures of Perrow's Work Unit Technology: An Empirical Assessment and a New Scale," *Academy of Management Journal* (26:1), pp. 45-63.
- Wolfers, J., and Zitzewitz, E. 2004. "Prediction Markets," *Journal of Economic Perspectives* (18:2), pp. 107-126.
- Zajonc, R. 1980. "Feeling and Thinking: Preferences Need No References," *American Psychologist* (35:2), pp. 151-175.

- Zigurs, I., and Buckland, B.K. 1998. "A Theory of Task/Technology Fit and Group Support Systems Effectiveness," *MIS Quarterly*.
- Zwass, V. 2010. "Co-Creation: Toward a Taxonomy and an Integrated Research Perspective," *International Journal of Electronic Commerce* (15:1), pp. 11-48.