Information Systems Research
Vol. 27, No. 1, March 2016, pp. 27–48
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2015.0605

Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing

Ivo Blohm
Institute of Information Management, University of St. Gallen, 9000 St. Gallen, Switzerland, ivo.blohm@unisg.ch

Christoph Riedl
D'Amore-McKim School of Business and College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; and Institute for Quantitative Social Science, Harvard University, Boston, Massachusetts 02138, c.riedl@neu.edu

Johann Füller
School of Management, University of Innsbruck, A-6020 Innsbruck, Austria, johann.fueller@uibk.ac.at

Jan Marco Leimeister
Chair for Information Systems, Kassel University, 34121 Kassel, Germany; and Institute of Information Management, University of St. Gallen, 9000 St. Gallen, Switzerland, leimeister@uni-kassel.de

Information technology (IT) has created new patterns of digitally-mediated collaboration that allow open sourcing of ideas for new products and services. These novel sociotechnical arrangements afford finely-grained manipulation of how tasks can be represented and have changed the way organizations ideate. In this paper, we investigate differences in behavioral decision-making resulting from IT-based support of open idea evaluation. We report results from a randomized experiment of 120 participants comparing IT-based decision-making support using a rating scale (representing a judgment task) and a preference market (representing a choice task). We find that the rating scale-based task invokes significantly higher perceived ease of use than the preference market-based task and that perceived ease of use mediates the effect of the task representation treatment on the users' decision quality. Furthermore, we find that the understandability of ideas being evaluated, which we assess through the ideas' readability, and the perception of the task's variability moderate the strength of this mediation effect, which becomes stronger with increasing perceived task variability and decreasing understandability of the ideas. We contribute to the literature by explaining how perceptual differences of task representations for open idea evaluation affect the decision quality of users and translate into differences in mechanism accuracy. These results enhance our understanding of how crowdsourcing as a novel mode of value creation may effectively complement traditional work structures.

Keywords: crowdsourcing; computer-mediated communication and collaboration; decision support systems; idea evaluation; rating scales; preference markets
History: Chris Forman, John Leslie King, Kalle Lyytinen, Senior Editors; Feng Zhu, Associate Editor. This paper was received June 30, 2013, and was with the authors 13 months for 3 revisions. Published online in *Articles in Advance* March 1, 2016.

1. Introduction

Digitization and improved information technology (IT) have had tremendous effects on the organization of work. Among the many changes, IT has been shown to affect decentralized work organizations (Tilson et al. 2010), knowledge work (Boudreau et al. 2014, Forman and Zeebroeck 2012), and decision-making (Arrow et al. 2008, Woolley et al. 2010). Many of these changes are driven by organizations' struggle to become more innovative and to satisfy the demands of heterogeneous users (Von Hippel 2005). To address these challenges, organizations increasingly experiment with open idea sourcing. IT changes the way organizations ideate. New patterns of digitally-mediated collaboration have emerged that facilitate open strategies of sourcing and the evaluation of innovative ideas for new products. The availability of open idea generation and evaluation mechanisms affects and changes the organization of work in companies that operate as sociotechnical systems (Lyytinen and King 2004). These mechanisms have allowed organizations to expand the generation and evaluation of ideas from a few select experts to the broader realms of the organization, possibly even reaching beyond the organization and involving customers (Afuah and Tucci 2012).

In addition to expanding the reach of who can contribute to idea generation and evaluation, modern IT-based decision-making support allows for the granular design of task representations. This is particularly

important in the area of idea evaluation, where we have a variety of evaluation strategies at our disposal. Two IT-based mechanisms are currently used in open idea evaluation: rating scales that represent idea evaluation as a judgment task (e.g., Di Gangi and Wasko 2009, Riedl et al. 2013) and preference markets that represent idea evaluation as a choice task (e.g., LaComb et al. 2007, Soukhoroukova et al. 2012). Yet representing a task as judgment- or choice-based brings with it several other critical differences in the IT-based implementation of such an evaluation mechanism. Judgment involves making decisions by individually assessing each alternative, whereas choice reflects the process of comparing among a set of alternatives and selecting those that are preferable (Moore 2004, Payne et al. 1992). Hence, a rating scale creates an absolute assessment set against the scale's endpoints and which has a meaningful interpretation by itself. Conversely, a preference market depends on a relative comparison of all contracts that represent the ideas to be evaluated. All contracts in the market must be known for a meaningful interpretation. Consequently, both task representations vary as to the interdependence of single idea evaluations (i.e., element interactivity) and the uncertainty in acquiring the information required to solve the decision task. Thus, the two task representations vary as to the cognitive effort required by their users and the resulting chance of making errors (Einhorn and Hogarth 1981). To our knowledge, these differences have not been systematically explored and organizations have not yet found satisfactory mechanisms for addressing the evaluation of early stage ideas and new opportunities (Chen et al. 2009). In this work, we investigate conditions under which task representations for open idea evaluation using rating scales and preference markets are similar or different and the general ease of use with which the corresponding IT-based task representations can be operated.

We report results of a Web experiment of idea evaluation with a random assignment of 120 participants to a rating scale or a preference market. We investigate differences between the two evaluation mechanisms related to (a) the evaluation task itself (which we assess through the readability of ideas), and (b) the task perception by the user through perceived ease of use and perceived task variability. We investigate these effects by drawing on the theories of cognitive load (Sweller 1988) and cognitive fit (Vessey and Dennis 1991).

Our analyses show that perceived ease of use mediates the effect of task representation on decision quality, such that the rating scale leads to a higher perceived ease of use which, in turn, results in higher decision quality. We also find that perceived task variability and readability of ideas moderate this mediation effect, such that users who perceive the evaluation task to be highly variable (i.e., perceive a higher cognitive burden) and evaluate difficult to understand ideas of low readability benefit more from increased perceived ease of use of the rating scale. Additionally, we show that users translate the increased decision quality of the rating scale into higher overall mechanism accuracy when aggregated across all users.

We make the following key contributions to understanding the parameters that improve the effectiveness of digital work arrangements for idea evaluation. First, task representation is crucial for the success of IT-based open idea evaluation (Afuah and Tucci 2012, Blohm et al. 2013, Estellés-Arolas and González-Ladrón-de-Guevara 2012, Zhao and Zhu 2014) but is poorly understood at this time (Dean et al. 2006, Ozer 2005, Riedl et al. 2013, Soukhoroukova et al. 2012). Our study explains differences in perceived ease of use resulting from the two different task representations. Second, we explain observed differences between idea evaluation based on rating scales and preference markets, taking into account the traits of the ideas being evaluated and the perception of the evaluation task; we also explore whether differences are consistent for ideas of low and high quality. Third, we contribute to crowdsourcing research by showing that task design and cognitive effort jointly affect decision quality in open idea evaluation.

2. Background

The systematic evaluation of early stage product ideas is of paramount importance for organizations (e.g., Girotra et al. 2010, LaComb et al. 2007, Ozer 2005, Riedl et al. 2013, Soukhoroukova et al. 2012). Open idea sourcing can only become valuable innovations if ideas can be evaluated in cost-efficient ways while also addressing ideas' increased diversity. Effective and efficient IT-based mechanisms for open idea evaluation promise to address two issues faced by organizations. First, they offer scalable approaches for addressing organizations' ongoing and expanding strategic identification of promising emerging opportunities (Chen et al. 2009). Scalable IT-based systems provide alternatives to costly expert panel evaluations, i.e., today's default mechanism for evaluation of early stage ideas (Hammedi 2011, Ozer 2005). Second, such effective IT-based mechanisms promise to improve the quality of idea evaluation by integrating diverse viewpoints and aggregating dispersed knowledge (e.g., Arrow et al. 2008, Riedl et al. 2013). Divisions of organizations often serve as standalone silos that can be broken up by IT-enabled decision-making support. Thus, well designed IT-based systems that provide effective representations of the evaluation task can help combat a leading factor in bad decision-making, i.e., the isolation of executives from the views and insights of the company's workforce (O'Leary 2013).

2.1. Measuring Idea Quality

Early stage ideas reflect a textual description of a *potential* solution to a problem of an organization. Idea quality consists of the dimensions of novelty, feasibility, relevance, and specificity (Dean et al. 2006). However, "true" idea quality is unknown until an idea has been implemented. When using open idea evaluation, the goal is to approach an idea's true quality, reflecting its expected value and assuming optimal allocation of resources during its implementation (Girotra et al. 2010). Thus, *mechanism accuracy* is a mechanism's ability to capture true idea quality to reduce a set of ideas to a subset (for further evaluation) or to select promising ideas for implementation. All ideation studies generally suffer from the fact that true quality is not observable (Chen et al. 2009, Girotra et al. 2010). Note that even if ideas are fully implemented and launched as products for which objective measures such as sales revenue could be collected, this would still not serve as an objective measure of idea quality, since the original idea likely would have been augmented during the implementation process (Boudreau et al. 2016, Kornish and Ulrich 2014). In the absence of objective measures of idea quality, we use a "best effort" alternative by developing a multisource measure of idea quality for increasing validity and reliability of this baseline for mechanism accuracy and users' decision quality. Our baseline is based on three independent expert and expert-user evaluations: (1) an expert assessment based on the consensual assessment technique (Amabile 1996), (2) an expert-based preference market approach (Dani et al. 2006), and (3) focus groups with expert-users (Morgan 1996).

2.2. Idea Evaluation, Mechanism Accuracy, and Decision Quality

Cognitive theory considers tasks at different levels of abstraction (Shaft and Vessey 2006). The task of evaluating ideas is represented by the mechanism that is provided to the user. Thus, mechanism accuracy is, all else being equal, a function of how well a mechanism supports decision-making of its users. The performance of such decision processes is generally referred to as *decision quality*, defined as the "correctness" or "goodness" of decisions (Dennis and Wixom 2001). It relates to the effectiveness with which users can solve a problem (Vessey and Dennis 1991). In the domain of evaluating innovation ideas, decision-making effectiveness refers to identifying the ideas of highest quality. Based on our multimethod estimation of true idea quality, we define decision quality as the effectiveness of mechanism use, i.e., the correctness of users' idea evaluations recorded with the mechanisms used. We apply this to idea evaluation and arrive at the levels of abstraction presented in Table 1.

Table 1 Abstraction Levels of an Idea Evaluation Task

Levels of abstraction	Idea evaluation task
Task	Evaluate idea quality
Representation	Use evaluation mechanism
Action	Record decisions about idea quality

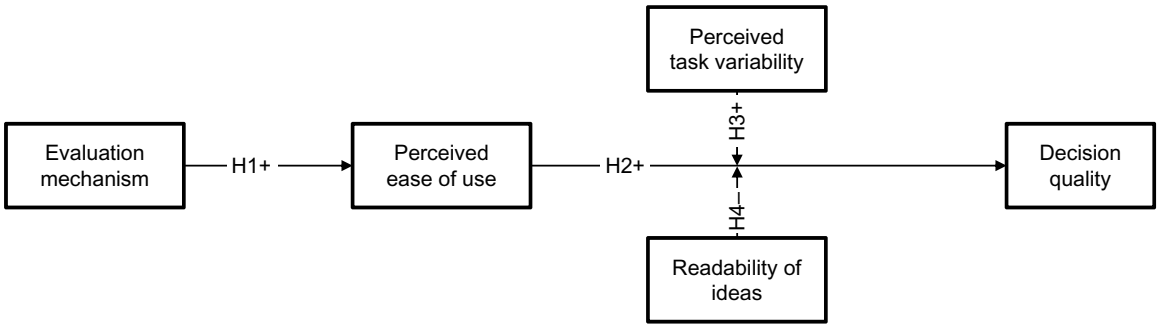
2.3. Idea Evaluation with Rating Scales

Behavioral decision-making research suggests that using a rating scale instantiates idea evaluation as a judgment task. Judgment involves making decisions about a set of alternatives in which each alternative is individually assessed. Using rating scales, users evaluate a finite set of alternatives (i.e., ideas) by applying a defined set of criteria. Assigning numerical values to these criteria, rating scales strive to identify an alternative that is closest to a defined optimum (Limayem and DeSanctis 2000). By means of different algorithms, individual ratings can be aggregated to group decisions (Todd and Benbasat 1999). The ideation literature has investigated rating scale-based procedures to evaluate the outcome of IT-enabled idea generation (e.g., Dean et al. 2006). However, this research has often focused on the use of rating scales by small groups of experts. For open idea evaluation, rating scales are frequently used (but rarely investigated) as a tool for eliciting evaluations from a larger group beyond select expert panels (e.g., Leimeister et al. 2009, Riedl et al. 2013, Zhao and Zhu 2014). Whereas rating scales by default do not provide peer feedback, typical applications for idea evaluation include some type of mean rating for users. In this regard, Di Gangi and Wasko (2009) as well as Blohm et al. (2013) have investigated how aggregated, rating-scale-based idea evaluations of customers affect organizational resource allocation decisions. In summary, there is a research gap in the current understanding of the circumstances in which rating scales can be effectively used for open idea evaluation.

2.4. Idea Evaluation on Preference Markets

In choice tasks, individuals make one holistic decision from which they select one choice, or a few choices from a larger set of alternatives. Preference markets are a frequently applied instantiation of idea evaluation as a choice task. Preference markets are a special type of prediction markets that are virtual markets for collecting, aggregating, and evaluating dispersed information (Soukhoroukova et al. 2012). In prediction markets, users buy contracts that are bound to future events, e.g., the outcomes of political elections or sporting events. These contracts then have a certain payoff if the event occurs; there is no payoff otherwise. In efficient prediction markets, the market price represents the market's expectation of the probability that the event will occur, and thus the market price acts as a predictor

Figure 1 Hypotheses and Research Model



for the event (Wolfers and Zitzewitz 2004). Through the market-based pricing mechanism, the market aggregates dispersed information of market participants (Arrow et al. 2008). The choice decision in such markets is implemented as selecting the contracts with the highest expected payoff. Since their first use in 1988, prediction markets have become widely used corporate applications (O'Leary 2013). Today, over 100 organizations have run internal markets, including companies such as General Electric (LaComb et al. 2007), Microsoft (Berg and Proebsting 2009), and Google (Cowgill and Zitzewitz 2015). Organizations frequently run markets for predicting sales or project deadlines (O'Leary 2013) in which predictive accuracy has frequently been high (Arrow et al. 2008).

A special corporate application reflects preference markets in which users trade contracts to assess the value of events for which no observable realization exists in the near future. Such markets, e.g., for idea evaluation, measure individual expectations about others' value perceptions or preferences for a given set of alternatives for which they create a price-based ranking (Chen et al. 2009, Dahan et al. 2010). As these markets cannot be tied to actual events, user payouts are commonly based on market-based measures such as price convergence (Chen et al. 2009) or artificial external events, e.g., expert evaluations (Soukhoroukova et al. 2012).

In our study, we used our multiple measurement of idea quality as the baseline, as we wanted to prevent market bubbles and "self-fulfilling prophecies" that can result from alternative market-based measures (Soukhoroukova et al. 2012). Most researchers explore preference markets as a novel approach to idea evaluation. LaComb et al. (2007) and Soukhoroukova et al. (2012) show how such markets can be used with employees, while Dahan et al. (2011) focus on use by customers. Trading behavior (Spears et al. 2009) or design choices such as contract design (Dahan et al. 2010), payout structures (Slamka et al. 2012) or incentives (Chen et al. 2009) have also been investigated. However, to our knowledge, it is as yet unanswered how the perception of the trading tasks that represent users' choice-based decisions affects users' ability to effectively operate the IT-based task representation, their ability to acquire information relevant to the choices they make, and the chance of making errors in their decision.

3. Theory Development and Hypotheses

Our study investigates decision processes in idea evaluation represented by a rating scale and a preference market. Drawing on cognitive theory, we investigate how differences related to the task itself and the task perception affect users' decision-making (Figure 1).

3.1. Perceived Ease of Use

Perceived ease of use is the degree to which a person believes that using a particular information system is free of effort (Davis 1989). It is directly rooted in using an information system and consists of the accompanying cognitive effort and ease with which users can learn to effectively use the information system (Davis 1989, Gefen and Straub 2000). While some traits of an idea evaluation task, such as the ideas to be evaluated, cannot be altered to support users in more effective decision-making, perceived ease of use is based on the characteristics of a task representation, i.e., an evaluation mechanism, and can be actively designed.

Cognitive load theory suggests that perceived ease of use is not only influenced by perceptions of tasks and their representations but also by a user's cognitive abilities (Sweller 1988). The basic assumption of cognitive load is that human cognition uses a short-term working and a long-term storage memory for processing information and tasks. The working memory processes all conscious cognitive tasks and can handle only a limited number of information elements (Van Merrienboer et al. 2003). Cognitive load theory distinguishes between cognitive load and cognitive effort.[1] Generally, cognitive load represents the load that performing a

[1] These dimensions are also frequently referred to as mental load and mental effort.

given task, or more specifically a task's representation, imposes on the cognitive system, i.e., the working memory of a user in terms of information processing demands (Chalmers 2003, Sweller et al. 1998). Cognitive effort refers to the amount of cognitive resources that a user must allocate to accommodate the task representation's information processing demands (Paas and Van Merriënboer 1994). Research has shown that the cognitive load associated with a task representation is best reflected by the cognitive effort that is needed for coping with that load (Paas et al. 2003). As perceived ease of using an evaluation mechanism is, all else being equal, driven by the cognitive effort imposed by a given evaluation mechanism, perceived ease of use can be conceptualized as the cognitive effort of using that evaluation mechanism (Chalmers 2003, Saadé and Otrakji 2007). Hence, the perceived ease of using a mechanism is grounded in two different dimensions of cognitive load imposed by this mechanism: intrinsic and extraneous (Kalyuga 2011, Paas et al. 2003).

Below we describe how aspects of rating scales and preference markets may affect perceived ease of use by affecting intrinsic and extraneous cognitive load. Intrinsic cognitive load refers to task-inherent complexity (Paas et al. 2003). While operating a rating scale is relatively intuitive, preference markets are quite complex (Kamp and Koen 2009). Cognitive load theory suggests that these differences are grounded in element interactivity that reflects the interconnectedness of information elements that must be considered in the working memory for using an evaluation mechanism (Kalyuga 2011, Sweller et al. 1998). While using a rating scale, idea evaluation exhibits a low level of element interactivity. By means of the rating scale, users judge the quality of each idea by the given evaluation criteria serially on a per idea basis. Users match the given criteria to existing subjective experiences and map them onto a numerical value on the scale (Limayem and DeSanctis 2000).

By contrast, the nature of choice inherent in preference markets exhibits a considerably higher level of element interactivity. The relative logic of preference markets forces users to compare ideas and to create one interrelated problem space. Users must integrate market prices of other idea contracts in the evaluation of one idea to make meaningful purchase or selling decisions. Furthermore, the evaluation decisions a user makes are interdependent, as choices are constrained by the user's liquid funds (Servan-Schreiber et al. 2004). These constraints may further increase the interconnectedness of elements that must be considered in the working memory during idea evaluation. In sum, adapting the flexible market logic to idea evaluation requires users to process more interacting elements, thus inducing higher intrinsic cognitive load compared to idea evaluation using a rating scale. This suggests that preference market users should face lower perceived ease of use than rating scale users.

Extraneous cognitive load refers to the way a task representation and the instructions for solving the task with that representation are organized. It arises due to changes in information architecture, visual complexity, media use, and information quantity (Jones et al. 2004, Mayer and Moreno 2003). In the domain of idea evaluation, a rating scale usually consists of one or more rating criteria to be judged, a visualization of the different response options (e.g., thumbs up/thumbs down icons or scales with a given number of categories), and some type of scale anchors for supporting users in interpreting the rating scale and decision-making (Riedl et al. 2013). These elements are usually presented in a spatially integrated fashion below the ideas that are evaluated. As a consequence, all information required for evaluating a given idea is provided, such that using a rating scale imposes little extraneous cognitive load (Mayer and Moreno 2003, Sweller et al. 1998).

Preference markets present a broad array of information such as market prices, price trends, portfolio value (i.e., the sum of liquid funds and the total value of hold idea contracts), and transaction histories to support users in decision-making and handling the market. Compared to rating scales, preference markets thus convey a considerable amount of additional information that dynamically changes over time due to the continuous interaction of market users. It is not only the resulting quantity of information processed by users that increases extraneous cognitive load, but since this information is also spatially separated on several parts of the preference market, it must be mentally integrated by users to make meaningful decisions. Consequently, users might require intense search processes and the recalling of some information elements while performing their transactions. Consequently, preference markets should invoke higher extraneous cognitive load than rating scales, such that preference market users should have lower perceptions of ease of use than rating scale users.

Generally, cognitive load imposed by a given task representation and the associated cognitive effort are additive, i.e., the total cognitive load an individual must process is the sum of the two different load types. In this regard, intrinsic cognitive load basically reflects the cognitive load that is directly grounded in a mechanism's basic logic and its way of working, whereas extraneous cognitive load may be derived from the mechanism's presentation to the user. Yet the total cognitive load that can be handled in the working memory also relates to an individual's prior experience (Paas et al. 2003). By frequently repeating similar tasks, users create cognitive structures (i.e., schemata) that allow them to bundle separate information elements into information chunks that can be

processed as one information element in the working memory. The acquisition of such cognitive structures enables experienced users to handle tasks of higher total cognitive load. In this vein, most individuals will already have had some exposure to using a rating scale, e.g., by responding to a survey or using evaluation functionalities on the Internet (Riedl et al. 2013). Thus, research on compatibility suggests that users can adapt these existing cognitive structures to using rating scales for idea evaluation. By contrast, preference markets are less common and most users will not have used them before, such that they would need to create new cognitive structures for using them (Karahanna et al. 2006). However, a task that requires creation of new cognitive structures is associated with a higher intrinsic cognitive load than a task for which existing cognitive structures can be adapted. Thus, rating scales should also invoke higher perceptions of ease of use than preference markets, as they are easier to learn.

Cognitive fit theory provides a complementary theoretical lens for explaining why a preference market may induce a higher cognitive load than a rating scale. The theory suggests that individuals develop separate mental representations for a task and its representation (Vessey and Dennis 1991). Cognitive fit reflects an interaction between both mental representations. In the case of good fit between a task and its representation, mental representations can be used straightaway to solve a given task; however, in the case of a poor fit, users must invest cognitive effort for sense-making and adapting mental models to solve the task, thus resulting in higher cognitive load (Shaft and Vessey 2006).

In evaluating ideas, users must develop mental representations of the evaluation task and the evaluation mechanism. Mental representations of the evaluation task require users to develop an understanding of the ideas and to form mental models of idea quality or preferences, depending on task representation. Mental representations of the mechanism reflect the logic of idea evaluation that a mechanism provides. In their exploratory study unraveling the cognitive processes of idea evaluation, Olshavsky and Spreng (1996) found that, independent of any mechanism or approach that facilitates an idea evaluation task, users judge ideas primarily by means of individually-formed evaluation criteria. As rating scales induce similar decision-making processes in which the given evaluation criteria frame the decision process (Riedl et al. 2013, Sujan 1985), we would expect high cognitive fit resulting in low cognitive load.

By contrast, adapting the relative choice-logic of preference markets, users must develop novel cognitive structures that should result in specific mental representations of the market. Adaptation of these mental representations to the task of idea evaluation poses an additional source of cognitive load. Hence, users must invest additional cognitive effort resulting in lower cognitive fit, compared to using a rating scale. Lower perceptions of ease of use should result.

In sum, both theories suggest that rating scales should be perceived to be easier to use than preference markets, as they impose a lower cognitive load and their use requires less cognitive effort. We thus purport:

Hypothesis 1 (H1). *The evaluation mechanism influences perceived ease of use, such that perceived ease of use will be higher for users of the rating scale than for users of the preference market.*

Both theories of cognitive load and cognitive fit suggest that perceived ease of use should positively influence the decision quality of users. Cognitive load theory assumes that cognitive load and cognitive effort are directly related to problem-solving effectiveness (i.e., decision quality). Within the limits of the capacity of their cognitive system, individuals can compensate for an increase of cognitive load by investing more cognitive effort to solve a given task, such that decision quality may remain constant even though cognitive load increases (Sweller et al. 1998). However, if the total cognitive load imposed by a task representation exceeds the capacity of an individual's cognitive system, cognitive overload is the consequence. In such circumstances, not all information elements necessary to solve the task can be processed in the working memory and thus the information processing capacity of the cognitive system is inhibited. Consequently, problem-solving effectiveness decreases, resulting in lower decision quality (Mayer and Moreno 2003, Todd and Benbasat 1999).

As perceived ease of use reflects the cognitive effort associated with using a given mechanism, it should be positively associated with decision quality. Low perceptions of ease of use can reflect situations in which users may have to invest high cognitive effort, suggesting that situations of cognitive overload may occur, which leads to lower decision quality. Similarly, high perceived ease of use reflects a state in which the total cognitive load imposed by an evaluation mechanism can be handled within the constrained working memory of the user. However, perceived ease of use reflects the cognitive effort of using a given evaluation mechanism and is thus directly grounded in the cognitive load imposed by the mechanism. Thus, cognitive load theory suggests that perceived ease of use should act as a mediator in the relationship between the task representation of an evaluation mechanism and decision quality.

Similarly, cognitive fit theory suggests that decision quality is highest when a given task fits its representation, as mental representations can be used straightaway to solve a given task (Vessey and Dennis 1991). As this may be the case for using rating scales, cognitive

fit theory suggests that using preference markets for idea evaluation requires adapting the mental representation of the task or its representation. As these mental operations reflect an additional source of cognitive load, higher cognitive effort is required to solve the task, leading to lower perceptions of ease of use. Consequently, the cognitive effort that can be allocated to the evaluation itself is limited, thus increasing the risk of cognitive overload. Higher levels of perceived ease of use should result in higher decision quality (Shaft and Vessey 2006, Vessey and Dennis 1991). Thus, we purport:

Hypothesis 2 (H2). *Perceived ease of use mediates the effect of the evaluation mechanism on decision quality, such that higher perceived ease of use leads to higher decision quality.*

3.2. Perceived Task Variability

Cognitive load theory suggests that cognitive capacity differs among individuals due to differences in intelligence and prior experiences through which cognitive structures for decision-making are formed (Paas and Van Merriënboer 1994, Sweller 1988). Thus, the cognitive capacity for handling tasks of varying cognitive load and the perception of task representations are subjective (Haerem and Rau 2007). Extending this argument, cognitive load theory suggests an interaction between cognitive capacity and task complexity. For instance, the results of Shaft and Vessey (2006) and Speier (2006) indicate that the cognitive load imposed by a task representation has a stronger negative effect on decision quality for complex tasks than it has for easy tasks. Their research suggests that the effect of perceived ease of use on decision quality might be moderated by perceived task variability.

Task variability is the number of exceptional cases, i.e., novel and unexpected stimuli, encountered while solving a task (Daft and Macintosh 1981). Perceived task variability, thus, describes how much a given task and its representation are nonroutine for the task solver and the associated feelings of surprise, uncertainty or difficulty (Haerem and Rau 2007). Generally, behavioral decision-making literature and cognitive load theory indicate that decision makers must acquire information to complete a decision task (Einhorn and Hogarth 1981, Sweller 1988). If a task is perceived as being variable, users must address a greater number of different information cues during information acquisition. Hence, high perceptions of task variability require users to invest additional cognitive effort into revealing the connections between the different and unexpected information cues during the phase of evaluation (Speier 2006). Thus, the cognitive load grounded in using an evaluation mechanism should be higher for users with high perceived task variability than for users with low perceived task variability (Paas and Van Merriënboer 1994).

Perceived task variability should then moderate the relationship between perceived ease of use and decision quality. For low perceived task variability, users should perceive the task as being more structured and more systematic with a lower number of unpredicted results. Users need invest little cognitive effort in handling unexpected stimuli and maintaining a sense of control while they use the mechanisms (Haerem and Rau 2007). Thus, there is small risk of cognitive overload hampering decision quality, as users perceiving the task to be less variable can operate the idea evaluation within the boundaries of their cognitive system (Shaft and Vessey 2006, Speier 2006). Consequently, the effect of perceived ease of use on decision quality should become weaker for users perceiving the task to be of low variability.

By contrast, users who perceive the task as highly variable may encounter many unexpected events while using a given evaluation mechanism (Haerem and Rau 2007). Handling such exceptions increases the cognitive load of mechanism use. Therefore users must increase cognitive effort to accommodate the task (Sweller et al. 1998). This increased cognitive load for users perceiving the task as highly variable may thus lead to more situations of cognitive overload in which users can no longer allocate the cognitive effort that is required to evaluate the ideas. Therefore, the positive effect of perceived ease of use on decision quality should be stronger for users who perceive the evaluation task as highly variable than for users who perceive the task to be less variable.

Given these arguments, perceived task variability should positively moderate the effect between perceived ease of use and decision quality. As perceived ease of use is invoked by the evaluation mechanism in the first place, perceived task variability may moderate the mediation between the evaluation mechanism and decision quality in which the strength of the mediating effect of perceived ease of use becomes stronger with increased levels of perceived task variability. We assume that:

Hypothesis 3 (H3). *Perceived task variability moderates the indirect effect of the evaluation mechanism on decision quality through perceived ease of use; the higher the perceived task variability, the greater the influence of perceived ease of use on decision quality.*

3.3. Readability of Ideas

The previous hypotheses relate to the perception of the evaluation task via a given evaluation mechanism. However, decision quality also depends on objective traits of the evaluation task beyond the evaluation mechanism that is representing the task. In this regard, salient traits of the ideas seem to be most relevant (Dean et al. 2006, Girotra et al. 2010). Intuitively, ideas that are easy to understand should be, all else being

equal, easier to evaluate than ideas that are more difficult to understand, as they impose less cognitive load. In a context in which evaluation objects are ideas represented through text, the readability of that text is critical, as readability has been defined as the ease of understanding or comprehension of text due to the style of writing (Klare 1963). Readability relates directly to the cognitive effort of understanding text (Ghose and Ipeirotis 2011). Existing research has shown that readability strongly influences human decision processes. For instance, Ghose and Ipeirotis (2011) show that higher readability of user-generated online reviews positively influences purchase decisions. Similarly, Tan et al. (2014) found that readability moderates the effect between the presentation of financial reports and the investors' earning evaluations.

Behavioral decision-making literature suggests that readability has a positive influence on information acquisition and evaluation in decision tasks (Einhorn and Hogarth 1981). When deriving information cues, textual idea descriptions are the primary source of information taken into account (Olshavsky and Spreng 1996). The more easily these textual idea descriptions and the information cues they provide can be understood, the less cognitive effort must be invested in distilling the relevant information for evaluation (Speier 2006). Thus, higher readability of textual idea representations should enable users to better understand idea traits and to build more precise mental representations of them.

This may be of great importance in online innovation communities in which ideas are user-generated content. In this context, ideas are often poorly written, contain spelling errors, and are difficult to understand (Blohm et al. 2011a, Ghose and Ipeirotis 2011). During the phase of information evaluation, blurry mental representations of an idea may lead users to ignore important information cues or to misinterpret them. The evaluation of such ideas imposes a higher cognitive load, and users must invest more cognitive effort for sense-making and addressing the increased uncertainty of information acquisition. Consequently, lower levels of readability should result in higher levels of intrinsic cognitive load (Ghose and Ipeirotis 2011).

Thus, readability should negatively moderate the effect of perceived ease of use on decision quality. As cognitive load is additive, the cognitive load grounded in evaluating ideas of varying readability will sum to the cognitive load of using a given evaluation mechanism that is represented by users' individual perceptions of ease of use (Mayer and Moreno 2003, Van Merrienboer et al. 2003). As evaluating ideas of low readability imposes a higher cognitive load than evaluating ideas of high readability, users who predominantly evaluate ideas of low readability should, all else being equal, face a higher total cognitive load than users who evaluate ideas of high readability. In instances in which users evaluate ideas of high readability, ideas are easy to understand, such that users can develop a clear mental representation of the ideas with comparably low cognitive effort and within the boundaries of their cognitive system. By contrast, users who predominantly evaluate ideas of low readability should face a higher risk of cognitive overload. Consequently, perceived ease of use should have a stronger effect on decision quality when users evaluate ideas of low readability compared to the evaluation of high readability ideas.

In sum, readability should negatively moderate the effect between perceived ease of use and decision quality. As we believe that perceived ease of use is grounded in the evaluation mechanism representing the idea evaluation task, readability may act as a moderator of the mediation between the evaluation mechanism and decision quality, such that the strength of the mediating effect of perceived task variability becomes weaker with increasing levels of readability. Thus, we posit:

Hypothesis 4 (H4). *Readability moderates the indirect effect of evaluation mechanism on decision quality through perceived ease of use; the higher the readability of the evaluated ideas, the smaller the influence of perceived ease of use on decision quality.*

4. Methodology

4.1. Experimental Task and Design

We designed and executed a Web experiment with a between-subject factorial design with two conditions, i.e., a preference market and a rating scale. We used a multicriteria rating scale composed of the criteria of novelty, value, feasibility, and specificity. The scale also contained an additional criterion with which users could indicate the confidence in their evaluations. Ideas were presented in a randomized order to rating scale users. Rating scale users updated their evaluations. The preference market was based on the Logarithmic Market Scoring Rules market maker (Hanson 2003) on which users of that treatment could repeatedly buy and sell contracts over a two-week trading period. Users could buy contracts of an idea if they believed that this idea would be among the five best ideas (i.e., "TOP-contracts"). The market also offered a down trading functionality in which users could indicate low perceived value. Specifically, users bought "FLOP-contracts" indicating that an idea would not be among the best five ideas. Users received a capital of 5,000 virtual currency units. They received a payoff of 100 virtual currency units for each idea contract they owned at the end of the market, and which was correctly classified, and 0 for incorrect classifications (Spann and Skiera 2003). We provide a detailed explanation of the preference market condition in the

online appendix (available as supplemental material at http://dx.doi.org/10.1287/isre.2015.0605). The configuration of the rating scale and of the preference market were identified in two extensive pretests of three different rating scales and six different preference market configurations, involving a total of 636 participants. Thus, they can be considered to be very robust.[2]

Users were randomly assigned to one of the two treatments. We invited users via a personalized email that included a link to the system URL and an online questionnaire, along with an exhaustive description of the task. We provided each user with a unique activation code to prevent cross-contamination and manipulations through the creation of multiple user accounts. Users completed the idea evaluation task distributed over the experiment duration of two weeks in November 2010 (Chen et al. 2009). Users had one week to complete the survey after the experiment.

We used a general platform for open idea sourcing developed by the authors. Features such as idea submissions and commenting were disabled. Only the evaluation mechanisms were activated. Apart from these mechanisms, both platforms were identical, consisting of a summary page containing the ideas, an overview page showing the ideas evaluated, and a FAQ section explaining the experimental task, and indicating how the mechanisms worked. For the preference market, the overview page also contained financial information, such as transaction prices, liquid funds, and a graph representing a user's overall value of all hold contracts. The system provided visual feedback for evaluations (e.g., highlighting rating scale buttons or updating price graphs) to make user interaction as easy as possible. Users participated with their own computers. As a Web experiment closely reflects the actual use scenarios of the mechanisms, high external validity of our results can be assumed. Users could evaluate the ideas in their natural environment and could allocate as much time as desired to complete the task.

4.2. Idea Sample

We used user-generated innovation ideas from a real-world community of the software producer SAP in Germany. This community invites SAP users to submit ideas to improve SAP software. The community evolved from an idea competition for students who had experience with SAP during their education (Leimeister et al. 2009). Today, the majority of the community's users are still students, and all ideas, except the one used in our experiment, were generated by students. We provide an idea example in the online appendix. At the time of the research, the community had collected 208 ideas, all of which were evaluated before the experiment by different experts. We aggregated these results, such that we drew a stratified sample of 24 ideas to reduce the workload of idea evaluation. Subsamples comprised eight ideas of high, medium, and low quality. The sample size was considered sufficient, as 20 to 30 ideas are used to measure decision quality of laypersons in creativity research (Runco and Smith 1992).

4.3. Participants

In our experiment, 132 users participated, of which 120 were included in the analysis. Users who did not complete the survey were removed. Participants were recruited from two large German universities. Our sample mainly consisted of Bachelor's and Master's students from two SAP-related management information systems courses. Furthermore, Ph.D. students in the field of information systems research took part in our experiment (all Ph.D. students had completed a Master's or equivalent). To incentivize participation, all Bachelor's and Master's students received homework credit points for participation. From the pool of all users, we also identified users with the highest decision quality in both treatment groups and raffled two MP3 players among them (similar to Slamka et al. 2012). Such payout schemes have been shown to enhance accuracy of prediction markets (Luckner and Weinhardt 2007). We applied multivariate analysis of variance to verify random assignment of users and found no systematic differences as to age, gender, and education between the treatments.

We follow Compeau et al. (2012) when discussing the appropriateness of our student sample. We define users participating in open idea sourcing as the target population. Such users tend to be male, young, and well educated (Franke and Shah 2003, Füller et al. 2009). A substantial fraction of them are students who are frequently motivated by the ability to signal their potential value to employers (Leimeister et al. 2009). For instance, Lakhani et al. (2013) report that 56% of users participating in a programming contest were students. In our experiment, 78.3% of our subjects were male, 19.2% had a Master's degree, 28.3% had a Bachelor's degree, and 52.5% finished high school. The mean age was 23.3 years.

The suitability of our user sample is also backed by its high concurrence with the users of the SAP community

[2] We tested the multicriteria scale against a thumbs up/thumbs down scale, and a five-star rating scale with 313 users. The multicriteria scale shows a higher rank-order correlation with an expert panel than the other two rating scales ($r = 0.47$). The applied preference market was pretested with a total of 323 users. In the preference market pretest, we varied the magnitude with which the market maker adapted prices after a single transaction and whether we provided users with a down trading option (see the online appendix for more details). The market used in this experiment had the strongest rank-order correlation with an expert panel ($r = 0.33$). Detailed results for these pretests have been reported in Riedl et al. (2010, 2013) and Blohm et al. (2011b).

that initially generated the ideas. We conducted the experiment at the end of the courses to ensure that our users had sufficient SAP-related knowledge to assess idea quality. Thus, we assume that idea generators (i.e., student users of the SAP community) did not have substantially more domain-specific knowledge than idea evaluators. Also, the setting of the evaluation task was created to be as realistic as possible (see §4.1). We can thus make a claim for parallelism between subject, task, and setting, suggesting that our users are appropriate subjects for our experiment (Compeau et al. 2012). Furthermore, research shows that students are appropriate for experiments that address the design of preference markets and rating scales (Luckner and Weinhardt 2007, Riedl et al. 2013, Slamka et al. 2012) and idea evaluation (Girotra et al. 2010).

5. Variables and Data Sources

We combined several data sources to investigate perceptual differences between the two task representation treatments: (1) behavioral use data of users' idea evaluations, (2) a quantitative survey of user cognitions, (3) a text-mining analysis of ideas' readability, and (4) three independent baseline measurements of decision quality. This data triangulation allows detailed insights into the complex interaction of user behavior, the cognitive perceptions of the idea evaluation task, and the investigated mechanisms as IT artifacts.

5.1. Decision Quality: Baseline Measurements of Idea Quality

To assess users' decision quality, we needed to construct an accurate measure for the a priori unknown, true quality of the ideas. Existing research shows that the uncertainty of innovation assessment can be reduced by combining multiple estimation methods (Ozer 2005). Drawing on the work of Campbell and Fiske (1959) and Girotra et al. (2010), we combined three independent measures of idea quality to improve the validity of our idea quality measurement. We applied two expert-based approaches for measuring idea quality: Amabile's (1996) consensual assessment technique and a preference market-based approach (Dani et al. 2006). Furthermore, we conducted five focus groups with experienced SAP users to measure idea quality (Morgan 1996). Thus, our baseline measures involve different evaluation methods as well as diverse judges. We compared these measurements of idea quality with the idea evaluations performed by the users to create an indicator for their decision quality. Below we describe the three baseline measures in more detail.

5.1.1. Baseline 1: Consensual Assessment Technique. The consensual assessment technique (CAT) was developed in creativity research for evaluating the quality of creative products and has already been used for user-generated ideas (e.g., Blohm et al. 2011a, Riedl et al. 2013). Our jury consisted of 11 referees who were professors of information systems, employees of SAP's marketing and research and development (R&D) department or the German SAP University Competence Centers. Idea quality was measured with four items used by SAP reflecting novelty, relevance, feasibility, and specificity. The ideas were copied into separate evaluation forms that were randomized and contained the scales for idea evaluation. Referees were assigned to rate the ideas independently with the four items on a five-point scale. We assessed the Intra-Class-Correlation-Coefficients (ICC) that should exceed 0.7 (Amabile 1996). We recognized this as being met for all items, excluding feasibility, of which the ICC was 0.5. Based on the mean quality of the ideas, we calculated a quality ranking.

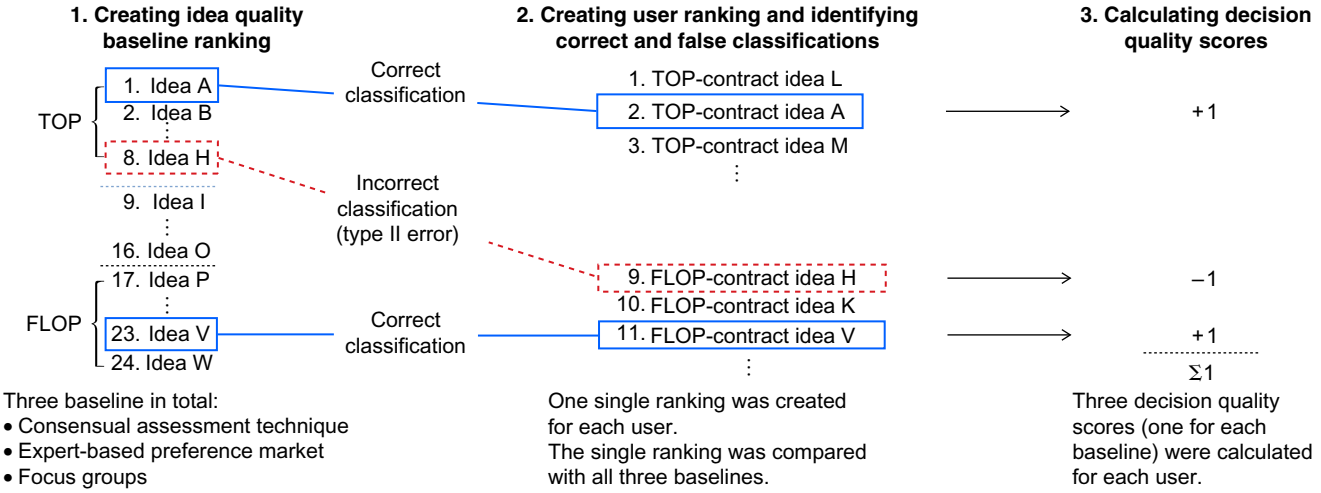
5.1.2. Baseline 2: Preference Market. Next, we measured idea quality based on a preference market-based procedure in which four experts from similar domains as used in the consensual assessment technique traded ideas on a preference market for one hour. Previous research has shown that prediction and preference markets can produce accurate results in such short trading periods in cases where all users use the market synchronously (Dahan et al. 2010, 2011). This is contrary to our experimental task in which users traded ideas asynchronously, making a longer trading period necessary (LaComb et al. 2007, Soukhoroukova et al. 2012). The preference market's configuration was adapted to the smaller number of participants. The experts were told to identify the five best ideas and had a seed capital of 5,000 virtual currency units. We used the market prices at the end of trading as a second ranking of idea quality.

5.1.3. Baseline 3: Focus Group Evaluations. We conducted five focus groups with four to six participants each. The 28 participants were drawn from an educational course for professional SAP users, reflecting a renowned SAP certification. Workshops typically lasted 75 minutes and consisted of three parts. First, participants received the ideas' descriptions in a randomized order and classified them into three groups of low, medium, and high quality. Second, participants collaboratively discussed the ideas' group assignments until they reached a consensus on which idea was part of which group. Finally, participants collaboratively ranked the ideas as to their quality in each of the three groups from which we obtained an overall quality ranking. To get a third ranking, we aggregated the five group rankings by arithmetic mean.

5.2. User Evaluations

In our experiment, users performed 1,507 transactions with the preference market and 4,471 ratings with the rating scale (across all criteria). These user evaluations were collected directly from the platform.

Figure 2 (Color online) Schematic Algorithm for Deriving Decision Quality Scores of Users



5.3. Triangulating User Evaluations with Baseline Evaluations

To assess an individual user's decision quality, we compared each user's evaluations to our baseline measures of idea quality and constructed a measure indicating each user's decision quality. In creativity research, decision quality of laypersons is usually determined by assessing the concurrent validity of their evaluations compared to an existing baseline such as experts, e.g., by counting correctly evaluated ideas (Runco and Smith 1992).

Figure 2 shows how we created our decision quality scores using the preference market treatment as an example. First, we created three baseline rankings reflecting the 24 ideas' quality according to the consensual assessment technique, the expert-based preference market, and the focus groups. In each ranking, we defined the best and worst eight ideas (33.3%) as "TOP"- and "FLOP"-ideas. We chose this cut off criterion, as about 10%–30% of user-generated ideas are typically of high quality (Blohm et al. 2011a).[3] Second, we created a quality ranking for each user based on her evaluations. For market users, the leading idea was indicated by the idea with the highest number of bought TOP-contracts. For rating scale users, the idea with the highest individual mean evaluation was leading. Comparing user rankings with each of the three baseline rankings, we identified correct and incorrect classifications by each user. In the market, we considered an idea to be correctly classified when users owned TOP-contracts of the eight TOP-ideas and FLOP-contracts of the eight FLOP-ideas at the market end. For rating scale users, we counted the best eight ideas that received a rating higher than the mean rating of each user and counted the worst eight ideas that received a rating lower than the mean. Finally, we calculated decision quality scores for each user by counting correctly classified ideas and subtracting classification errors (TOP-ideas classified as FLOP and vice versa). Applying this procedure, we rendered three decision quality measures for each user.

A major trait of prediction and preference markets is that users weigh their evaluations in terms of confidence represented by the number of contracts bought or sold within a single transaction (Spann and Skiera 2003). As we believe this information is important to describe preference market users' ability to evaluate idea quality and as similar information is not available using rating scales, we asked rating scale users to evaluate the confidence in each of their evaluations with a five-point rating scale. To integrate the two different confidence measurements, we computed confidence weights for the obtained decision quality scores. The confidence weights are based on an idea evaluation's deviation from the mean number of bought contracts or from the mean confidence rating. For preference market users, we calculated for each idea not only the mean number of contracts that were held by all users at the market end but also the corresponding standard deviations. For each idea, we created three reference groups by adding (subtracting) a half standard deviation to (from) the idea's mean amount. We compared each idea evaluation with the three reference groups to obtain a confidence weight. For idea evaluations in the low confidence group, decision quality scores were multiplied by 1, in the moderate confidence group by 2, and in the high confidence group by 3. For rating scale users, we followed the same procedure but used confidence ratings for creating the three reference groups and making the comparisons.

[3] In various robustness tests we confirm that our results are robust against variations of this cut-off criterion.

5.4. Perceived Ease of Use and Perceived Task Variability: Questionnaire

Data on perceived ease of use and perceived task variability were collected with a post-experimental online questionnaire. We used established scales to measure perceived ease of use (Davis 1989, Gefen and Straub 2000) and perceived task variability (Haerem and Rau 2007). All items were measured on a five-point scale. The survey was pretested with a sample of 10 users, reflecting the different user groups. Based on this feedback, we made minor changes.

5.5. Evaluation Mechanism: Dummy Coding

We operationalized the evaluation mechanisms as a dummy variable in which the preference market served as the reference group (preference market = 0; rating scale = 1).

5.6. Readability: Text Mining

We performed basic text mining analysis of the textual description of ideas to assess ideas' readability. Today, there are numerous readability metrics, and while none is perfect, they correlate well with the actual difficulty of reading a text. To avoid idiosyncratic errors peculiar to a specific metric, we followed the same approach as that of Ghose and Ipeirotis (2011) and computed different metrics.[4] In our main analyses, we report only results using the Coleman–Liau index; results are robust to alternative measures. The readability measure captures the average length of words and the average length of sentences, thus also capturing some aspects of content complexity. For each user, we aggregated readability as the median of the ideas that the user evaluated to account for indivisibility of ideas. The Coleman–Liau index, like most other readability measures, reports results as school grade level equivalents, such that lower numeric values indicate higher readability (DuBay 2004). To align the dimensionality of the readability construct and its measurement, we transformed the obtained index values, such that higher values indicate higher readability. For better interpretability, we also transformed the readability scores ranging from 0 (lowest) to 1 (highest).

5.7. Control Variables: Behavioral User Data

We included a time control as well as the number of idea evaluations to capture the level of participation.

[4] Specifically, we used the koRpus package (Michalke 2015) for the *R* language and environment for statistical computing to compute the Automated Readability Index (ARI), the Coleman–Liau Index, the Flesch–Kincaid Grade Level, the Gunning Frequency of Gobbledygook (FOG), the Simple Measure of Gobbledygook (SMOG), and Andersen's Readability Index. DuBay (2004) provides a detailed description on how to compute these metrics. The online appendix provides descriptive statistics and correlations for these metrics.

6. Results

Our research model implies a moderated mediation effect that generally describes *when* and *under what* conditions an effect occurs, i.e., that the strength of a mediation effect is based on a moderator (Preacher et al. 2007). A variable is a mediator when it represents the generative mechanism through which one variable influences another. More specifically, a focal independent variable influences the mediator that, in turn, impacts a dependent variable in a causal fashion (Baron and Kenny 1986, Shrout and Bolger 2002). By contrast, moderators influence strength and direction of a relationship between two variables. Consequently, moderated mediation models reflect mediation models in which one or several relationships in a mediation model are moderated (Hayes 2013, Preacher et al. 2007).

For testing the moderated mediation effect, we applied ordinary least squares (OLS) regressions with a nonparametric bootstrapping approach to compute bias-corrected confidence intervals (Hayes 2013, Preacher et al. 2007). Generally, mediation occurs when the magnitude of a direct effect between an independent and a dependent variable is weakened when a mediator variable is introduced in that relationship. Thus, the applied procedure involves a direct bootstrapping-based mediation test. The test looks for a significant difference between the strength of the direct effect between an independent and a dependent variable and the effect between the two variables, which is controlled for the mediator.[5] Bootstrapping shows low type I errors and high power in assessing moderated mediation effects (Preacher et al. 2007); it also has higher predictive validity than alternatives such as the causal steps (Baron and Kenny 1986) or the product of coefficient approach (Sobel test) for testing mediation (Shrout and Bolger 2002).

6.1. Construct Validation

To confirm validity and reliability of our measures, we applied exploratory and confirmatory factor analysis using SPSS 19 and SmartPLS 2.0 (see the online appendix). The Measure of Sampling Adequacy was 0.7,

[5] This bootstrapping procedure conceptualizes a study sample (of size N) as a pseudo-population from which the study sample was derived. Randomly drawing N samples with replacement from this population, a point estimate for the difference of the direct effect without mediator and the direct effect while controlling for mediator is calculated (Hayes 2013, Preacher et al. 2007). Repeating this procedure for a desired number of bootstrapping resamples, confidence intervals can be constructed. For hypothesis testing, the null hypothesis of no mediation effect is rejected at the desired level of significance if 0 lies outside the confidence interval (Preacher et al. 2007). For testing moderated mediation, confidence intervals that correspond to the different values of the moderator are calculated, i.e., the N samples in each bootstrapping resample are only drawn from a subpopulation in which the moderator variable satisfies a desired level.

Table 2 Descriptive Statistics

	Rating scale users			Preference market users			All users		
	Mean (SD)	Min.	Max.	Mean (SD)	Min.	Max.	Mean (SD)	Min.	Max.
(1) DQ	5.93 (6.88)	−8.33	28.33	2.31 (3.54)	−7.00	9.00	4.31 (5.89)	−8.33	28.33
(2) PEOU	4.14 (0.70)	2.33	5.00	3.67 (0.75)	1.67	5.00	3.93 (0.76)	1.67	5.00
(3) PTV	2.76 (0.83)	1.00	4.67	2.51 (0.83)	1.00	4.00	2.65 (0.83)	1.00	4.67
(4) Readability	0.76 (0.06)	0.00	1.00	0.72 (0.16)	0.00	1.00	0.74 (0.12)	0.00	1.00
(5) Time (min)	78.85 (138.83)	1.23	849.23	171.01 (249.32)	2.67	1,221.17	120.32 (200.77)	1.23	1,221.17
(6) Evaluations (#)	16.97 (7.09)	1.00	24.00	27.91 (25.63)	3.00	164.00	21.89 (18.70)	1.00	164.00

Notes. SD, Standard deviations; DQ, decision quality; PEOU, perceived ease of use; PTV, perceived task variability.

indicating good applicability of exploratory factor analysis. We used the scree test criterion for extracting four factors that could be clearly interpreted. The scree test was used, as the latent root criterion (Eigenvalues > 1) tends to extract too few factors for factor analysis with fewer than 20 items, such as in our study (Hair et al. 2010). Applying the latent root criterion, our single item for readability (of high importance for our study) would have been deleted, as it has an Eigenvalue < 1. Alphas of at least 0.7 suggest good reliability of factors. Composite Reliabilities (CR) exceeded values of 0.5, and the Average Variance Explained (AVE) for each factor surpassed 0.5. Thus, convergent validity could be assumed (Bagozzi and Yi 1988). However, we eliminated one item from our perceived task variability scale due to a nonsignificant factor loading. The discriminant validity was checked using the Fornell–Larcker criterion, which claims that the one factor's AVE should be higher than its squared correlation with every other factor (Fornell and Larcker 1981). Thus, discriminant validity could be assumed.

6.2. Descriptive Statistics and Correlations

Before testing the hypotheses, we present some basic descriptive statistics. Tables 2 and 3 depict means, standard deviations, minimum and maximum values, as well as correlations of our study variables. Rating scale users exhibit higher decision quality ($p < 0.01$) and higher perceived ease of use ($p < 0.01$). Evaluation behavior differed among mechanisms. Rating scale users made 16.97 evaluations on average, while preference markets made an average of 27.91 evaluations. The difference in the number of evaluations is significant ($p < 0.01$) and results from the fact that preference market users traded each idea contract about 2.5 times, whereas rating scale users did not similarly update their ratings. Overall, preference market users spent at least 90 minutes longer on the experimental platform ($p < 0.01$). To control for these behavioral differences, we included time and the number of evaluations as control variables in our user-level analysis.

Table 3 Correlations

	(1)	(2)	(3)	(4)	(5)
(1) DQ					
(2) PEOU	0.25**				
(3) PTV	0.12	−0.20*			
(4) Readability	0.06	0.20*	−0.04		
(5) Time	−0.11	−0.06	0.05	−0.14	
(6) Evaluations	−0.03	−0.04	0.13	0.04	0.46**

Notes. $N = 120$; DQ, decision quality; PEOU, perceived ease of use; PTV, perceived task variability.
*$p < 0.05$; **$p < 0.01$.

6.3. Regression and Bootstrapping Analysis

As interaction effects, including moderation and mediation, may come in many different forms, we followed Muller et al. (2005), who suggest testing moderated mediation with a series of OLS regressions. This approach allows us to investigate our research model, while ruling out alternative models. We applied the PROCESS SPSS Macro with 1,000 bootstrapping intervals to assess the significance of the moderated mediation effect (Hayes 2013). We followed Cohen et al. (2003) and used z-standardized SPSS factor scores (decision quality, perceived ease of use, and perceived task variability) and indicators (readability and controls). The treatment variable was not z-standardized. We estimated three regressions.

$$\begin{aligned} &\textit{Decision Quality} \\ &\quad = \alpha + \beta_1\, \textit{Rating Scale} + \beta_2\, \textit{Perceived Task Variability} \\ &\qquad + \beta_3\, \textit{Rating Scale} \times \textit{Perceived Task Variability} \\ &\qquad + \beta_4\, \textit{Readability} + \beta_5\, \textit{Rating Scale} \times \textit{Readability} \\ &\qquad + \beta_6\, \textit{Time} + \beta_7\, \textit{Evaluations} + \textit{error}; \end{aligned} \tag{1}$$

Table 4 Regression Results for Decision Quality

		Equation (1)	Equation (2)	Equation (3)
Tested effect	Independent variable	DV: Decision quality	DV: Perceived ease of use	DV: Decision quality
T	Rating scale	0.33**	0.34**	0.27**
MO	PTV	−0.14	−0.21	−0.03
$T \times$ MO	Rating scale × PTV	0.26	−0.03	0.19
MO	Readability	0.02	0.09	−0.11
$T \times$ MO	Rating scale × Readability	−0.01	0.15	−0.14
ME	PEOU			0.27**
ME × MO	PEOU × PTV			0.23*
ME × MO	PEOU × Readability			−0.15
	Time	−0.07	0.00	−0.08
	Evaluations	0.14	0.10	0.11
R^2		0.14*	0.20**	0.24**

Notes. $N = 120$; T, Treatment; MO, moderator; $T \times$ MO, treatment × moderator; ME, mediator; ME × MO, mediator × moderator; DV, dependent variable; PEOU, perceived ease of use; PTV, perceived task variability.
*$p < 0.05$; **$p < 0.01$.

$$\begin{aligned}\textit{Perceived Ease of Use} &= \alpha + \beta_1\, \textit{Rating Scale} + \beta_2\, \textit{Perceived Task Variability} \\ &\quad + \beta_3\, \textit{Rating Scale} \times \textit{Perceived Task Variability} \\ &\quad + \beta_4\, \textit{Readability} + \beta_5\, \textit{Rating Scale} \times \textit{Readability} \\ &\quad + \beta_6\, \textit{Time} + \beta_7\, \textit{Evaluations} + \textit{error}; \end{aligned} \tag{2}$$

$$\begin{aligned}\textit{Decision Quality} &= \alpha + \beta_1\, \textit{Rating Scale} + \beta_2\, \textit{Perceived Task Variability} \\ &\quad + \beta_3\, \textit{Rating Scale} \times \textit{Perceived Task Variability} \\ &\quad + \beta_4\, \textit{Readability} + \beta_5\, \textit{Rating Scale} \times \textit{Readability} \\ &\quad + \beta_6\, \textit{Perceived Ease of Use} \\ &\quad + \beta_7\, \textit{Perceived Ease of Use} \times \textit{Perceived Task Variability} \\ &\quad + \beta_8\, \textit{Perceived Ease of Use} \times \textit{Readability} + \beta_9\, \textit{Time} \\ &\quad + \beta_{10}\, \textit{Evaluations} + \textit{error}. \end{aligned} \tag{3}$$

Equation (1) establishes a direct effect of the rating scale treatment on decision quality, including perceived task variability and readability as moderators. Equation (2) reflects the same regression equation but uses perceived ease of use as a dependent variable. Equation (3) tests the effect of our treatment on decision quality while controlling for perceived ease of use as a mediator, as well as taking into account the effects of the moderators (perceived task variability and readability). To test whether the residual effect of the treatment is influenced by the assumed moderators, we also considered respective interaction terms (i.e., rating scale × perceived task variability and rating scale × readability) (Muller et al. 2005). Based on these equations, the bootstrapping procedure applies a direct test for moderated mediation, indicating whether perceived ease of use is the generative mechanism through which the task representation influences decision quality. We report regression results in Table 4. Bootstrapped confidence intervals are shown in the online appendix.

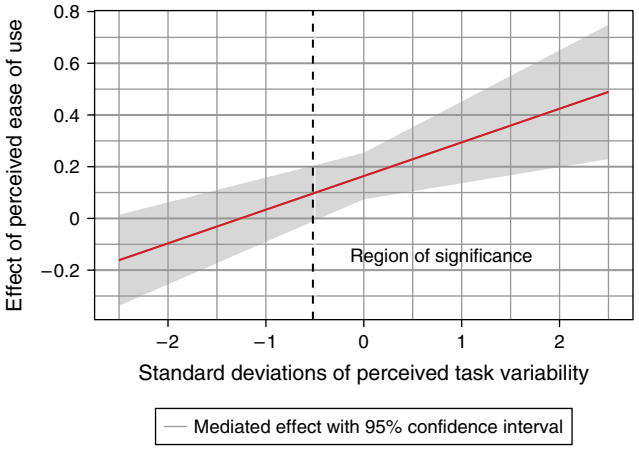
Testing Equation (1), our results suggest that the rating scale treatment directly affects decision quality ($\beta = 0.33$; $p < 0.01$), indicating that the rating scale leads to a higher decision quality than the market. Neither perceived task variability nor readability have a statistically significant moderating effect on decision quality, indicating that the rating scale-decision quality path in our model is not moderated.[6] Testing Equation (2), we find that our treatment influences perceived ease of use, indicating that the rating scale is perceived to be easier to use than is the market ($\beta = 0.34$; $p < 0.01$). Furthermore, we find that the positive effect of using a rating scale on decision quality is not moderated by perceived task variability or readability.

Next, we investigate the mediating role of perceived ease of use estimating Equation (3). We find that perceived ease of use is positively associated with decision quality ($\beta = 0.27$; $p < 0.01$). Furthermore, we find that perceived task variability positively moderates the effect of perceived ease of use on decision quality ($\beta = 0.23$; $p < 0.05$). However, we find that readability does not have a significant effect on the perceived ease of use, decision quality relationship ($\beta = -0.15$; $p =$ not significant). We must assess the bootstrapped confidence intervals to test moderated mediation. Bootstrapping indicates that the mediation effect of perceived ease of use is significant when users perceive moderate and high levels of task variability, while the evaluated ideas are of low and moderate readability ($p < 0.05$). Given these results, the mediating role of perceived ease of use is stronger when users perceive the evaluation task to be highly variable. Similarly, this effect becomes weaker when users evaluate ideas of lower readability.

Both the OLS regression-based procedure (Muller et al. 2005) and the bootstrapped moderated mediation

[6] Beyond the suggestions of Muller et al. (2005), we also verified that perceived ease of use does not moderate the evaluation mechanism, i.e., the decision quality relationship.

Figure 3 (Color online) Strength of Mediation Effect of Perceived Ease of Use by Perceived Task Variability



test (Hayes 2013) show support for H1–H3. As to H4, the effect of readability on the perceived ease of use–decision quality relation results are mixed. Bootstrapping suggests a significant moderation effect of readability, while OLS regression analysis does not. However, as existing research points to the superior predictive validity of bootstrapping in testing combined indirect effects (Hayes 2013, Preacher et al. 2007, Shrout and Bolger 2002), we accept H4.

All models control for observed behavioral differences in mechanism use, which are not significantly correlated with decision quality (Equations (1) and (3)) or perceived ease of use (Equation (2)). Furthermore, we verified the robustness of our results, testing our models with alternative operationalizations of decision quality. We ran separate analyses for each of the three decision quality indicators and alternative aggregations of them (arithmetic mean, geometric mean, median) as well as without weighing the decision quality scores for confidence. All analyses are consistent with the main results presented here.

Figure 4 (Color online) Strength of Mediation Effect of Perceived Ease of Use by Readability

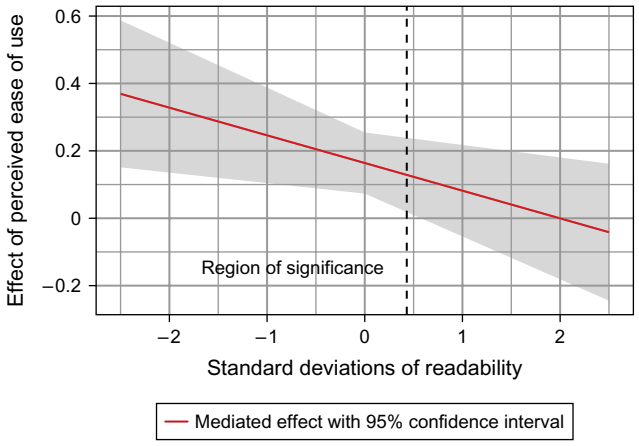
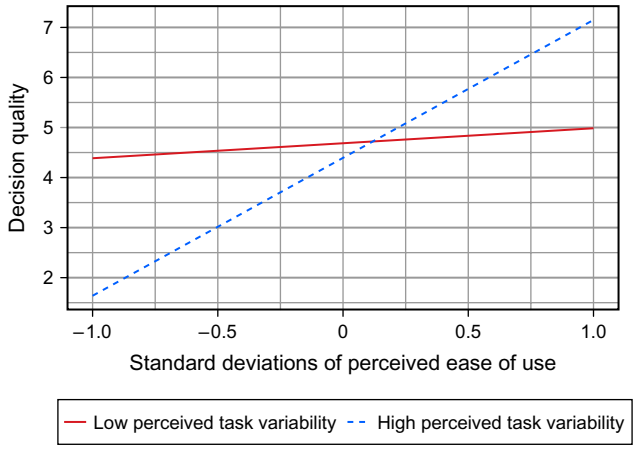


Figure 5 (Color online) Marginal Means for Decision Quality and Perceived Task Variability Interaction

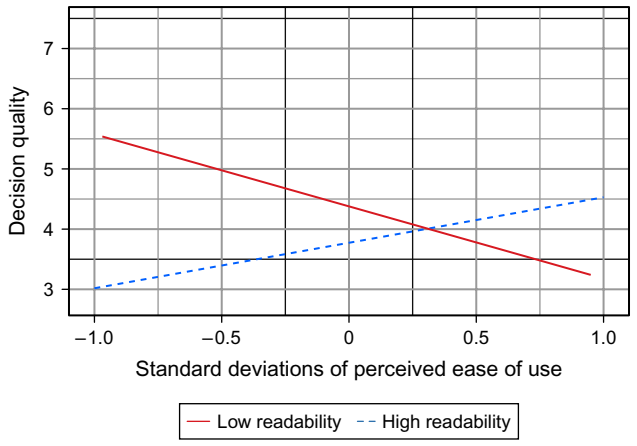


We probe our moderation analyses through visual representations. For plotting the moderated mediation effect (see Figures 3 and 4), we calculated the strength of the mediation effect of perceived ease of use at different levels of the moderators, including the threshold value for which the mediating effect is significant (Preacher et al. 2007). Results indicate that the mediation effect is significant for values of perceived task variability that are at least −0.37 standard deviations above the mean and for values of readability that are at least 0.39 standard deviations smaller than the mean. Figures 5 and 6 show the interaction between perceived ease of use and perceived task variability as well as readability. These plots suggest that easy-to-use mechanisms are particularly important in settings of high perceived task variability and low readability.

6.4. Sensitivity Analysis: Evaluation of FLOP- and TOP-Ideas

A key factor that can affect the function of any evaluation mechanism is the characteristic of the decision-making

Figure 6 (Color online) Marginal Means for Decision Quality and Readability Interaction



task itself (Olshavsky and Spreng 1996, Sujan 1985). In the context of idea evaluation, such a key characteristic could be whether a user is evaluating ideas of high or low quality. To investigate if differences in the decision-making task affect how the two evaluation mechanisms used in this study function, we separately analyzed data for the evaluation of FLOP- and TOP-ideas. If evaluating ideas of, say, low quality, was a simpler task than evaluating ideas of high quality, we might expect differences between how perceived ease of use affects the use of an evaluation mechanism. This difference in difficulty is relevant, as research has shown that individuals often approach complex decision-making situations differently from less complex situations (Payne et al. 1992, Todd and Benbasat 1999). First, we perform a simple comparison of mean values of users' decision quality scores. For rating scale users, we find a mean decision quality of 1.69 for TOP-ideas and a mean decision quality of 4.24 for FLOP-ideas ($p < 0.01$). For preference market users, we find a mean decision quality of 1.01 for TOP-ideas and a mean decision quality of 1.31 for FLOP-ideas ($p =$ not significant). Second, we re-tested Equations (1) and (3) (Equation (2) remains unchanged) using the correct classification of only FLOP- or TOP-ideas as dependent variables (Table 5; see the online appendix for bootstrapping results). Results of this analysis are largely consistent with the main analyses presented above but with two marked differences: For FLOP-ideas the mediation effect is only significantly moderated by perceived task variability; for TOP-ideas we find no direct effect of the evaluation mechanism treatment. We discuss each of these differences in turn.

The bootstrapping analysis suggests that when evaluating FLOP-ideas the mediation effect of perceived ease of use does not depend on the level of readability. This suggests that decision quality does not significantly degrade in more complex decision-making instances in which users evaluate ideas that are hard to read. This indicates that evaluating FLOP-ideas is a simpler task compared to evaluating ideas of all quality levels, and that users' decision quality does not deteriorate even in difficult situations.

By contrast, the evaluation of TOP-ideas seems to be more difficult. A key difference suggested by the analysis is that when evaluating TOP-ideas, we find no statistically significant difference between users of the rating scale and the preference market. Also, assessing bootstrapping intervals offers no clear picture. Compared with the analysis that does not distinguish between FLOP- and TOP-ideas, the difference in decision quality between the users of the two mechanisms becomes lower when considering only the evaluation of TOP-ideas. This is explained by the fact that decision quality of the preference market users does not degrade as drastically when performing the more complex task of evaluating high-quality ideas compared to the decision quality degradation faced by rating scale users. To investigate this difference, we compared the user's mean decision quality scores in the set of TOP-ideas compared to the set of FLOP-ideas. For rating scale users, decision quality is 60.11% lower when rating TOP-ideas compared to FLOP-ideas. This is a statistically significant difference. In other words, users make many more mistakes when evaluating TOP-ideas compared to FLOP-ideas. Preference market users also face lower decision quality evaluating TOP-ideas compared to FLOP-ideas which is 23.10% lower compared to when evaluating FLOP-ideas. This is not a statistically significant difference. Thus, users of preference markets

Table 5 Regression Results for Evaluating FLOP- and TOP-Ideas

		Classification of FLOP-ideas		Classification of TOP-ideas	
		Equation (1)	Equation (3)	Equation (1)	Equation (3)
Tested effect	Independent variable	DV: Decision quality[a]	DV: Decision quality[a]	DV: Decision quality[b]	DV: Decision quality[b]
T	Rating scale	0.39*	0.30**	0.12**	0.13
MO	PTV	−0.05	0.06	−0.18	−0.10
$T \times$ MO	Rating scale × PTV	0.06	0.02	0.36*	0.29*
MO	Readability	0.12	0.06	−0.09	0.24
$T \times$ MO	Rating scale × Readability	−0.02	−0.13	−0.01	−0.09
ME	PEOU		0.36**		0.06
ME × MO	PEOU × PTV		0.17*		0.19
ME × MO	PEOU × Readability		0.03		−0.21
	Time control	−0.01	−0.02	−0.11	−0.10
	Evaluations	0.09	0.07	0.13	0.11
R^2		0.17**	0.28**	0.09	0.16*

Notes. $N = 120$; T, Treatment; MO, moderator; $T \times$ MO, treatment × moderator; ME, mediator; ME × MO, mediator × moderator; DV, dependent variable; PEOU, perceived ease of use; PTV, perceived task variability.

[a]Correct classification of TOP-ideas (correctly classified TOP-ideas subtracted by error from incorrectly classified TOP-ideas (Type II error)).

[b]Correct classification of FLOP-ideas (correctly classified FLOP-ideas subtracted by error from incorrectly classified FLOP-ideas (Type I error)).

*$p < 0.05$; **$p < 0.01$.

Table 6 Mechanism Accuracy

	MAPE (lower is better)	Haessel R^2 (higher is better)	DQ score[a] (higher is better)	DQ score[a] TOP-ideas	DQ score[a] FLOP-ideas
RS accuracy	0.65	0.52	11	5	6
PM accuracy	0.75	0.50	11	5	6
N	24	24	24	8	8

Notes. Ex-post fit between rankings produced by rating scale and preference market and "true" ranking produced by expert baseline measured by MAPE, Haessel R^2, and DQ score. MAPE, Mean absolute percentage error; DQ, decision quality; RS, rating scale; PM, preference market.
[a]Same decision quality score as applied in the user level analysis.

perform comparatively better on the complex task of evaluating TOP-ideas. Question: What could explain this difference?

One possible explanation could be derived from differences in how individuals approach complex decisions. Prior research suggests that individuals try to reduce complexity by basing their decisions on the decisions of others (Easley and Kleinberg 2010) and that they particularly follow positive information cascades (Muchnik et al. 2013). Faced with complex decisions, individuals look for additional information cues that would help them make decisions. Because the preference market provides users with signals of how others have assessed the situation through the information aggregation of the market price, this gives them an opportunity to base their own decisions on those of others. As incorporating such information cues within the evaluation of TOP-ideas corresponds to following positive price trends (i.e., positive information cascades), the information aggregation property of the preference market may support users in the more complex task of correctly identifying TOP-ideas. The rating scale, however, does not provide such signals and users cannot reduce decision-making complexity by basing their own decisions on those of others.[7] This positive aspect of supporting users' decision-making in the more difficult decision-making situation of evaluating TOP-ideas thus seems to offset some of the negative effects faced by preference markets, thereby evening out the performance discrepancy between preference markets and rating scales.

7. Supplementary Mechanism-Level Results

Previous analyses of the user level reported on differences in perceived ease of use between representation of the idea evaluation task as rating scale or preference market. While the main focus of our analysis has been investigating cognitive mechanisms and individual-level effects of decision-making task representations on decision quality, it is natural to wonder if differences observed on the user level would translate into any differences at the mechanism level when aggregating across users to arrive at collective decisions. Hence, we performed additional supplementary analyses on the aggregated level to investigate collective decision-making outcomes. To increase the robustness of our results, we performed the analyses using three separate measures: the decision quality score as applied on the user level, as well as the Mean Absolute Percent Error (MAPE) and Haessel R^2, which are commonly used in prediction market research to evaluate forecasting errors (Goodwin and Lawton 1999, King et al. 1993).[8] For rating scales, we aggregated all user ratings for each idea by geometric mean, which is the recommended aggregation in collective intelligence tasks (Lorenz et al. 2011). For the expert baseline, we also aggregated the ranks by geometric mean, and for prediction market rankings we used final market prices (Table 6). Aggregating by arithmetic mean does not substantially change our results.

Comparing totals by treatment, we calculate percentage increase in fit between a rating scale and a preference market and our true expert baseline by using MAPE as $100(0.75-0.65)/0.65 = 14.83\%$. Using Haessel R^2, the increase in ex-post fit with the expert baseline when comparing the rating scale and the preference market is 4.80%. Using the decision quality

[7] We find some support for this possibility in our preference market data. We followed Chen et al. (2009) and classified socially influenced transactions (i.e., transactions that support a positive or negative price trend when compared to the two previous transactions). Evaluations of TOP-ideas exhibit stronger social influence than evaluations of FLOP-ideas ($p < 0.1$). In a regression in which positively and negatively biased transactions are regressed on the correct classification of TOP-ideas, following positive price trends has a considerable positive effect on correctly classifying TOP-ideas ($\beta = 0.19$; $p < 0.01$) while following negative price trends creates a small classification error ($\beta = -0.07$; $p < 0.01$). When evaluating FLOP-ideas, following negative price trends ($\beta = -0.16$; $p < 0.01$) creates little advantage compared to the error of following positive price trends ($\beta = -0.15$; $p < 0.01$). Thus, the social signals of a preference market particularly support users in evaluating TOP-ideas.

[8] In these analyses, we used the ranks predicted by our experiment (forecast ranking) and the baseline rankings (actual ranking) to analyze mechanism accuracy.

Table 7 Leave-One-Out Cross Validation

	MAPE (lower is better)	Haessel R^2 (higher is better)	DQ score[a] (higher is better)	DQ score[a] TOP-ideas	DQ score[a] FLOP-ideas
RS accuracy mean	0.69	0.50	10.79	5.04	6.00
PM accuracy mean	0.87	0.35	8.33	3.54	5.00
Difference in mean	−0.18**	0.14**	2.46**	1.50**	0.96**
N	23	23	23	7–8	7–8

Notes. MAPE, Mean absolute percentage error; DQ, decision quality; RS, rating scale; PM, preference market.
[a]Same decision quality score as applied in the user level analysis.
**$p < 0.01$.

score, the rating scale and the preference market are identical. To test whether these differences in mechanism accuracy are statistically significant, we performed leave-one-out cross validation and then performed a paired *t*-test (Witten et al. 2011). Table 7 shows that the rating scale produces significantly lower ranking error than the market ($p < 0.01$), higher Haessel R^2 ($p < 0.01$), and a significantly higher decision quality score ($p < 0.1$). Furthermore, we tested for statistical difference in classification error between TOP- and FLOP-ideas. For the rating scale mechanism, the mean difference in decision quality is −0.71 ($p < 0.01$) and for the preference market, the mean difference is −1.25 ($p < 0.01$). This indicates that correctly identifying TOP ideas is a significantly harder task that can only be completed with higher errors. In summary, across the three different measures, we find significantly higher accuracy with the rating scale.

8. Discussion and Implications

Innovation is an important and constantly challenging business activity. Changes in the nature of work as a result of using IT and a shift to open strategies for organizing work and sourcing ideas have changed the ways in which innovation activities are organized. While various methods for open idea generation have achieved considerable maturity, idea evaluation remains a challenge. Along with trends to source ideas openly, new IT-enabled evaluation mechanisms have been proposed that rely on fundamentally different representations of the idea evaluation task. In this paper, we analyze perceptual differences between two leading mechanisms for open idea evaluation that implement the evaluation task using rating scales and preference markets.

8.1. Theoretical Implications

We show that representing idea evaluation as a rating scale invokes higher perceived ease of use than a preference market and that these differences significantly influence decision quality of users. The mediating role of perceived ease of use between the rating scale treatment and decision quality is strengthened when considering perceived task variability and readability as moderators. To our knowledge, we are the first to perform these analyses in the scope of open idea evaluation. We expand our understanding of perceptual differences among IT-based implementations for digitally-mediated knowledge work and coordination, making three main contributions.

First, we go beyond existing studies in the ideation literature by grounding rating scales and preference markets in behavioral decision theory by offering a deeper understanding of these mechanisms for open idea evaluation, investigating the perceptual differences of the task representations that these mechanisms create, and relating these perceptual differences to decision quality of users. We attempt to integrate two streams of research in the ideation literature that have separately studied the aggregation of distributed information about idea quality. Existing studies of preference markets for idea evaluation have addressed applying these markets in various real-life settings (LaComb et al. 2007, Soukhoroukova et al. 2012), considered high-level design choices such as payout structures (Slamka et al. 2012) and incentives (Chen et al. 2009) or investigated user behavior (Spears et al. 2009). Similarly, rating scales for idea evaluation have been studied as a tool for collecting idea evaluations to create idea rankings (Di Gangi and Wasko 2009, Riedl et al. 2013). We answer the call of various researchers for a more detailed study of collective decision-making tools that are now commonly used in open idea evaluation (Kamp and Koen 2009, LaComb et al. 2007, Soukhoroukova et al. 2012).

Second, we contribute to theory in the area of crowdsourcing by extending our knowledge of *how* crowdsourcing and participatory systems may effectively complement traditional work arrangements of organizations (Afuah and Tucci 2012, Estellés-Arolas and González-Ladrón-de-Guevara 2012, Zhao and Zhu 2014). We find that the effect of task representations and consequent evaluation mechanism implementations on users' decision quality is mediated by perceived ease of use. We can thus support the notion that better usability of crowdsourcing mechanisms should lead to higher levels of decision quality. This is important, as ease of use of information systems can be systematically influenced through design decisions. Our results

suggest that rating scales invoke higher perceived ease of use, resulting in higher decision quality. These findings complement research on aggregating individual opinions in the domain of crowdsourcing by showing that mechanism accuracy is not only a function of the aggregation mechanism, as frequently assumed in research on prediction (e.g., Arrow et al. 2008, Hanson 2003, Spann and Skiera 2003) and preference markets (e.g., Dahan et al. 2010, LaComb et al. 2007), but is also reflected in the decisions formed on the user level, which may be systematically hampered by evaluation mechanisms that are too difficult to use.

Third, our findings contribute to the research field of behavioral decision theory (e.g., Moore 2004, Nowlis and Simonson 1997, Payne et al. 1992). We investigate novel representations of judgment and choice tasks that are increasingly instantiated by means of IT (e.g., Kühberger and Gradl 2013, Moore 2004). To our knowledge, in this field, traits and perceptions of IT-enabled decision mechanisms have not yet been systematically addressed. Our results show that designing easy-to-use mechanisms may help support users in their decision-making, and could be applied in various settings, e.g., web-based conjoint analyses to make more precise preference assessments or to explain performance differences.

8.2. Practical Implications

Our results help mitigate the risks of overly complex evaluation mechanisms by making recommendations for the design of idea evaluation tasks and their representation as evaluation mechanisms. First, our results indicate that presenting idea evaluation by a rating scale can lead to better decision quality of users and higher mechanism accuracy than instantiating idea evaluation by a preference market. These results are particularly strong when evaluating FLOP-ideas, which seems to be an easier task. If the goal of an idea evaluation task is to reduce the size of a set of ideas, the task should be framed so that users are asked to filter out low quality ideas. This is contrary to the current practice in which users are typically asked to collectively identify high quality ideas. Second, our results point to the pivotal role of ease of use. Designing easy-to-use evaluation mechanisms is paramount for effective open idea evaluation processes, pointing to the role of systematic usability testing. Ease of use frees cognitive resources and allows users to make more accurate idea evaluation decisions. Similarly, evaluation mechanisms should be designed so that they structure the idea evaluation task and provide a high sense of control to the users, as unexpected stimuli during idea evaluation undermine the decision quality of users. These design choices are particularly important when ideas of high complexity are evaluated, as low readability increases cognitive load and overall task complexity.

These recommendations must be viewed in light of the relation between the costs of idea evaluation and evaluation errors. These results are of high practical relevance, as the impact of Type I errors (i.e., wrongly classifying a FLOP-idea as TOP) and Type II errors (i.e., wrongly classifying a TOP-idea as FLOP) in innovation may be different. While implementing ideas in the former case simply reflects a misallocation of resources, the latter may reflect a lost opportunity that could be fatal to an organization.

8.3. Limitations and Future Research

There are some limitations to our study. Potential limitations may arise as to external validity from the use of students and the specific focus of our idea sample on enterprise software. However, as our experiment establishes a high concurrence between users, task, and setting, our results should be generalizable to most other settings in which ideas represented as text are evaluated by novice users (Compeau et al. 2012). Results may differ for image-based representations processed via different cognitive mechanisms. Our user cognition measures, such as for perceived ease of use, were based on a survey. Future research might use verbal protocols (Olshavsky and Spreng 1996) or neurophysiological tools such as functional Magnetic Resonance Imaging (fMRI) (Dimoka et al. 2012), which are more comprehensive and objective methods. These methods could give rise to a deeper understanding of individual decision-making processes and accompanying task characteristics such as cognitive load (Dimoka et al. 2012). While these methods may reflect a fruitful avenue for future research, research on cognitive load has shown that individuals can transform their perceived effort into a numerical value. Thus, survey-based measures of user cognitions provide reasonably reliable and valid data while not interfering with the decision task and providing high sensitivity (Paas et al. 2003).

Preference markets for idea evaluation generally suffer from the fact that no real observable outcomes exist. Users could be betting on expected expert evaluations and not on idea quality itself. Thus, more research is needed on how preference markets can be operated for which no real observable outcomes exist (Slamka et al. 2012). However, the long-term effectiveness of preference market and open idea evaluation in general has still to be studied to improve their use in digitally-mediated work settings (Cowgill and Zitzewitz 2015).

In our experiment, the two mechanisms representing the idea evaluation task were implemented in their most basic fashion, turning off various functionalities that allow social interaction (beyond the market mechanism). In practice, platforms for open idea sourcing tend to be highly interactive to spur communication and collaboration among participants. While this was a deliberate decision based on prior work (Hildebrand et al. 2013),

it may lead to a slight deviation from real world-based settings, in which users may also provide qualitative feedback, i.e., comments that may dynamically change ideas. We found it more important to focus on our main condition of interest, i.e., the cognitive mechanism explaining performance differences between behavioral decision-making resulting from IT-based support of open idea evaluation, without introducing additional confounding effects. Future research could extend our research by explicitly adding experimental conditions to study the interaction of cognitive decision processes and social interaction.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/isre.2015.0605.

Acknowledgments

The authors thank the senior editors, the associate editor, and the reviewers for their guidance and valuable feedback during the review process. The authors also thank Markus Rieger, who provided software implementation help, and Bernd Skiera, who provided valuable feedback on an earlier draft. The authors would like to thank Sabine Matook, Andrew Burton-Jones, Iris Vessey, and the research workshop participants at the Australian National University, University of Queensland for their valuable inputs and helpful comments. The first author received support from the basic research fund of the University of St. Gallen.

References

Afuah A, Tucci C (2012) Crowdsourcing as a solution to distant search. *Acad. Management Rev.* 37(3):355–375.

Amabile TM (1996) *Creativity in Context. Update to Social Psychology of Creativity* (Westview Press, Boulder, CO).

Arrow KJ, Forsythe R, Gorham M, Hahn R, Ledyard JO, Levmore S, Litan R, et al. (2008) The promise of prediction markets. *Sci.* 320(5878):877–878.

Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. *J. Acad. Marketing Sci.* 16(1):74–94.

Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Personality Soc. Psych.* 51(6): 1173–1182.

Berg H, Proebsting TA (2009) Hanson's automated market maker. *J. Prediction Markets* 3(1):45–59.

Blohm I, Leimeister JM, Krcmar H (2013) Crowdsourcing: How to benefit from (too) many great ideas. *MIS Quart. Executive* 12(4):199–211.

Blohm I, Bretschneider U, Leimeister JM, Krcmar H (2011a) Does collaboration among participants lead to better ideas in IT-based idea competitions? An empirical investigation. *Internat. J. Networking Virtual Organ.* 9(2):106–122.

Blohm I, Riedl C, Leimeister JM, Krcmar H (2011b) Idea evaluation mechanisms for collective intelligence in open innovation communities: Do traders outperform raters? *Proc. Internat. Conf. Inform. Systems (ICIS'11), Shanghai.*

Boudreau K, Guinan E, Lakhani K, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance and resource allocation in science. *Management Sci.* ePub ahead of print January 8, http://dx.doi.org/10.1287/mnsc.2015.2285.

Boudreau K, Gaule P, Lakhani KR, Riedl C, Woolley AW (2014) From crowds to collaborators: Initiating effort and catalyzing interactions among online creative workers. Working Paper 14-060, Harvard Business School, Cambridge, MA.

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psych. Bull.* 56(2): 81–105.

Chalmers PA (2003) The role of cognitive theory in human-computer interface. *Comput. Human Behav.* 19(5):593–607.

Chen L, Goes P, Marsden JR, Zhang Z (2009) Design and use of preference markets for evaluation of early stage technologies. *J. Management Inform. Systems* 26(3):45–70.

Cohen J, Cohen P, West SG, Aiken LS (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Mahwah, NJ).

Compeau D, Marcolin BL, Kelley H, Higgins C (2012) Generalizability of information systems research using students—A reflection on our practices and recommendations for future research. *Inform. Systems Res.* 23(4):1093–1109.

Cowgill B, Zitzewitz E (2015) Corporate prediction markets: Evidence from Google, Ford, and Firm X. *Rev. Econom. Stud.* 86(4): 1309–1341.

Daft RL, Macintosh NB (1981) A tentative exploration into the amount and equivocality of information processing in organizational work units. *Admin. Sci. Quart.* 26(2):207–224.

Dahan E, Soukhoroukova A, Spann M (2010) New product development 2.0: Preference markets how scalable securities markets identify winning product concepts and attributes. *J. Product Innovation Management* 27(2):937–954.

Dahan E, Kim AJ, Lo AW, Poggio T, Chan N (2011) Securities trading of concepts (STOC). *J. Marketing Res.* 48(3):497–517.

Dani V, Madani O, Pennock D, Sanghai S, Galebach B (2006) An empirical comparison of algorithms for aggregating expert predictions. Dechter R, Richardson T, eds. *Proc. 22nd Conf. Uncertainty Artificial Intelligence (UAI)* (AUAI Press, Arlington, VA), 106–113.

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* 13(3):319–340.

Dean DL, Hender JM, Rodgers TL, Santanen EL (2006) Identifying quality, novel, and creative ideas: Constructs and scales for idea evaluation. *J. Assoc. Inform. Systems* 7(10):646–698.

Dennis AR, Wixom BH (2001) Investigating the moderators of the group support systems use with meta-analysis. *J. Management Inform. Systems* 18(3):235–257.

Di Gangi PM, Wasko MM (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):303–312.

Dimoka A, Banker RD, Benbasat I, Davis FD, Dennis AR, Gefen D, Gupta A, et al. (2012) On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS. *MIS Quart.* 36(3):679–702.

DuBay WH (2004) *The Principles of Readability* (Impact Information, Costa Mesa, CA).

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

Einhorn HJ, Hogarth RM (1981) Behavioral decision theory: Processes of judgment and choice. *J. Accounting Res.* 19(1):1–31.

Estellés-Arolas E, González-Ladrón-de-Guevara F (2012) Towards an integrated crowdsourcing definition. *J. Inform. Sci.* 38(2):189–200.

Forman C, Zeebroeck NV (2012) From wires to partners: How the Internet has fostered R&D collaborations within firms. *Management Sci.* 58(8):1549–1568.

Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. *J. Marketing Res.* 18(2):39–50.

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

Füller J, Mühlbacher H, Matzler K, Jawecki G (2009) Consumer empowerment through Internet-based co-creation. *J. Management Inform. Systems* 26(3):71–102.

Gefen D, Straub D (2000) The relative importance of perceived ease of use in IS adoption: A study of e-commerce adoption. *J. Assoc. Inform. Systems* 1(8):1–28.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowl. Data Engrg.* 23(10):1498–1512.

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management Sci.* 56(4):591–605.

Goodwin P, Lawton R (1999) On the asymmetry of the symmetric MAPE. *Internat. J. Forecasting* 15(4):405–408.

Haerem T, Rau D (2007) The influence of degree of expertise and objective task complexity on perceived task complexity and performance. *J. Appl. Psych.* 92(5):1320–1331.

Hair JF, Black WC, Babin BJ, Anderson RE (2010) *Multivariate Data Analysis* (Pearson, Boston).

Hanson R (2003) Combinatorial information market design. *Inform. Systems Frontiers* 5(1):107–119.

Hammedi W, van Riel ACR, Sasovova Z (2011) Antecedents and consequences of reflexivity in new product idea screening. *J. Product Innovation Management* 28(5):662–679.

Hayes AF (2013) *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Guilford Press, New York).

Hildebrand C, Häubl G, Herrmann A, Landwehr JR (2013) When social media can be bad for you: Community feedback stifles consumer creativity and reduces satisfaction with self-designed products. *Inform. Systems Res.* 24(1):14–29.

Jones Q, Ravid G, Rafaeli S (2004) Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration. *Inform. Systems Res.* 15(2): 194–210.

Kalyuga S (2011) Cognitive load theory: How many types of load does it really need? *Ed. Psych. Rev.* 23(1):1–19.

Kamp G, Koen PA (2009) Improving the idea screening process within organizations using prediction markets: A theoretical perspective. *J. Prediction Markets* 3(2):39–64.

Karahanna E, Agarwal R, Angst CM (2006) Reconceptualizing compatibility beliefs in technology acceptance research. *MIS Quart.* 30(4):781–804.

King R, Smith V, Williams A, van Boening M (1993) The robustness of bubbles and crashes in experimental stock markets. Day RH, Chen P, eds. *Nonlinear Dynamics and Evolutionary Economics* (Oxford, New York), 183–200.

Klare GR (1963) *Measurement of Readability* (Iowa State University Press, Ames, IA).

Kornish LJ, Ulrich KT (2014) The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *J. Marketing Res.* 51(1):14–26.

Kühberger A, Gradl P (2013) Choice, rating, and ranking: Framing effects with different response modes. *J. Behav. Decision Making* 26(2):109–117.

LaComb AC, Barnett A, Pan Q (2007) The imagination market. *Inform. Systems Frontiers* 9(2–3):245–256.

Lakhani KR, Boudreau KJ, Loh P-R, Backstrom L, Baldwin C, Lonstein E, Lydon M, MacCormack A, Arnaout RA, Guinan EC (2013) Prize-based contests can provide solutions to computational biology problems. *Nature Biotechnology* 31(7):108–111.

Leimeister JM, Huber M, Bretschneider U, Krcmar H (2009) Leveraging crowdsourcing—Activation-supporting components for IT-based idea competitions. *J. Management Inform. Systems* 26(1):197–224.

Limayem M, DeSanctis G (2000) Providing decisional guidance for multicriteria decision making in groups. *Inform. Systems Res.* 11(4):386–401.

Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc. Nat. Acad. Sci.* 108(22):9020–9025.

Luckner S, Weinhardt C (2007) How to pay traders in information markets: Results from a field experiment. *J. Prediction Markets* 1(2):147–156.

Lyytinen K, King JL (2004) Nothing at the center?: Academic legitimacy in the information systems field. *J. Assoc. Inform. Systems* 5(6):220–247.

Mayer RE, Moreno R (2003) Nine ways to reduce cognitive load in multimedia learning. *Ed. Psych.* 38(1):43–52.

Michalke ME (2015) *koRpus: An R package for text analysis*. Accessed February 17, 2016, https://cran.r-project.org/web/packages/koRpus/index.html.

Moore WL (2004) A cross-validity comparison of rating-based and choice-based conjoint analysis models. *Internat. J. Res. Marketing* 21(3):299–312.

Morgan DL (1996) *Focus Groups as Qualitative Research* (Sage, Thousands Oaks, CA).

Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Sci.* 341(6146):647–651.

Muller D, Judd CM, Yzerbyt VY (2005) When moderation is mediated and mediation is moderated. *J. Personality Soc. Psych.* 89(6): 852–863.

Nowlis SM, Simonson I (1997) Attribute-task compatibility as a determinant of consumer preference reversals. *J. Marketing Res.* 34(2):205–218.

O'Leary DE (2013) Internal corporate prediction markets: "From each according to his bet." *Internat. J. Accounting Inform. Systems* 14(1):89–103.

Olshavsky RW, Spreng RA (1996) An exploratory study of the innovation evaluation process. *J. Product Innovation Management* 13(6):512–529.

Ozer M (2005) Factors which influence decision making in new product evaluation. *Eur. J. Oper. Res.* 163(3):784–801.

Paas F, Tuovinen J, Tabers H, van Gerven P (2003) Cognitive load measurement as a means to advance cognitive load theory. *Ed. Psych.* 38(1):63–71.

Paas FG, Van Merriënboer JJ (1994) Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *J. Ed. Psych.* 86(1):122–133.

Payne JW, Bettman JR, Coupey E, Johnson EJ (1992) A constructive process view of decision making: Multiple strategies in judgment and choice. *Acta Psych.* 80(1–3):107–141.

Preacher KJ, Rucker DD, Hayes AF (2007) Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behav. Res.* 42(1):185–227.

Riedl C, Blohm I, Leimeister JM, Krcmar H (2010) Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right. *Proc. Internat. Conf. Inform. Systems (ICIS'10), St. Louis*, 1–21.

Riedl C, Blohm I, Leimeister JM, Krcmar H (2013) The effect of rating scales on decision quality and user attitudes in online innovation communities. *Internat. J. Electronic Commerce* 17(3):7–37.

Runco MA, Smith WR (1992) Interpersonal and intrapersonal evaluations of creative ideas. *Personality Individual Differences* 13(3): 295–302.

Saadé GR, Otrakji CA (2007) First impressions last a lifetime: Effect of interface type on disorientation and cognitive load. *Comput. Human Behav.* 23(1):525–535.

Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B (2004) Prediction markets: Does money matter? *Electronic Markets* 14(3): 243–251.

Shaft TM, Vessey I (2006) The role of cognitive fit in the relationship between software comprehension and modification. *MIS Quart.* 30(1):29–55.

Shrout PE, Bolger N (2002) Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psych. Methods* 7(4):422–445.

Slamka C, Jank W, Skiera B (2012) Second-generation prediction markets for information aggregation: A comparison of payoff mechanisms. *J. Forecasting* 31(1):469–489.

Soukhoroukova A, Spann M, Skiera B (2012) Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets. *J. Product Innovation Management* 29(1):100–112.

Spann M, Skiera B (2003) Internet-based virtual stock markets for business forecasting. *Management Sci.* 49(10):1310–1326.

Spears B, LaComb C, Interrante J, Barnett J, Senturk-Dogonaksoy D (2009) Examining trader behavior in idea markets: An implementation of GE's imagination markets. *J. Prediction Markets* 3(1):17–39.

Speier C (2006) The influence of information presentation formats on complex task decision-making performance. *Internat. J. Human-Comput. Stud.* 64(11):1115–1131.

Sujan M (1985) Consumer knowledge: Effects on evaluation strategies mediating consumer judgments. *J. Consumer Res.* 12(1):31–46.

Sweller J (1988) Cognitive load during problem solving: Effects on learning. *Cognitive Sci.* 12(2):257–285.

Sweller J, van Merrienboer J, Paas F (1998) Cognitive architecture and instructional design. *Ed. Psych. Rev.* 10(3):251–296.

Tan H-T, Ying Wang E, Zhou BO (2014) When the use of positive language backfires: The joint effect of tone, readability, and investor sophistication on earnings judgments. *J. Accounting Res.* 52(1):273–302.

Tilson D, Lyytinen K, Sörensen C (2010) Digital infrastructures: The missing IS research agenda. *Inform. Systems Res.* 21(4):748–759.

Todd P, Benbasat I (1999) Evaluating the impact of DSS, cognitive effort, and incentives on strategy selection. *Inform. Systems Res.* 10(4):356–374.

Van Merrienboer JJG, Kirschner PA, Kester L (2003) Taking the load off a learner's mind: Instructional design for complex learning. *Ed. Psych.* 38(1):5–13.

Vessey I, Dennis G (1991) Cognitive fit: An empirical study of information acquisition. *Inform. Systems Res.* 2(1):63–84.

Von Hippel E (2005) *Democratizing Innovation* (MIT Press, Cambridge, MA).

Witten IH, Frank E, Hall MA (2011) *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, Burlington, MA).

Wolfers J, Zitzewitz E (2004) Prediction markets. *J. Econom. Perspective* 18(2):107–126.

Woolley A, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Sci.* 330(6004):686–688.

Zhao Y, Zhu Q (2014) Evaluation on crowdsourcing research: Current status and future direction. *Inform. Systems Frontiers* 16(3): 417–434.