<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *The Journal of product innovation management*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Magnusson, P R., Wästlund, E., Netz, J. (2016)

Exploring users' appropriateness as a proxy for experts when screening new product/service ideas.

The Journal of product innovation management, 33(1): 4-18

<https://doi.org/10.1111/jpim.12251>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-33920>

Exploring users' appropriateness as a proxy for experts when screening new product/service ideas

Peter R. Magnusson, Associate Professor

Erik Wästlund, Associate Professor

Johan Netz, PhD Candidate

All authors are affiliated to the Service Research Center, Karlstad University, Sweden

Corresponding author is Peter Magnusson e-mail: peter.magnusson@kau.se

Biography

Dr. Peter R. Magnusson, is Associate Professor at the Service Research Center at Karlstad University, Sweden. He holds an MSc in electrical engineering from Chalmers University, an MBA in executive business administration from the University of Uppsala, and a Ph.D. from the Stockholm School of Economics. Dr. Magnusson has twenty years' practical experience in R&D in the computing and telecommunications industries. His research focuses on new product and service innovation, with a special interest for open innovation and user involvement. Dr. Magnusson has received several nominations and rewards for his research. He has been published in leading refereed journals, including *Journal of Product Innovation Management*, *Journal of Service Research*, *Journal of the Academy of Marketing Science*, and *Creativity and Innovation Management*.

Dr. Erik Wästlund is a senior lecturer in psychology at Karlstad University. Since he received his PhD from the University of Gothenburg (2007) his main research focus has been on consumer behavior and decision making. Erik has been working with eye-tracking in a number of different R & D contexts such as marketing, human-computer-interaction and assistive technology.

Johan Netz is a Ph.D. student at Karlstad University. His research focuses on new product and service innovation, with special interest for idea management and open innovation. In addition to research he also lecture in marketing strategy and user-innovation.

Magnusson, P. R., E. Wästlund, and J. Netz. 2016. Exploring users' appropriateness as a proxy for experts when screening new product/service ideas. *Journal of Product Innovation Management*, 33 (1): 4-18.

Abstract

Despite the popularity and extensive use of engaging users in crowdvoting, very little research has been conducted into the appropriateness of users as substitutes for experts when judging ideas. Accordingly, the purpose of this article is to investigate the appropriateness of using users as a proxy for professional experts during the initial idea screening of new product/service ideas. In other words, are users' assessments during idea screening conformant with professional experts' assessments and are they reliable as a proxy for experts during idea screening? In a comparative study, two different approaches to outsourcing the screening of wireless ideas to users are examined, including assessment by two different user panels: (1) Technically-Skilled Users, (2), and Technically-Naïve Users. These two approaches were compared with the assessments made by professional experts. The results showed no conformance between users and experts when comparing their absolute scores. However, during a relative comparison (the ranking of ideas), both user panels were conformant with the professional experts. A test of the user panels' ability to select the same top ideas as the professional experts was successful, indicating good conformance between the user panels and the professional experts. This article's contribution is knowledge of how conformant external users are compared with professional expert judges during idea screening. The results indicate that companies can employ users during the initial screening process using criteria assessment to select the best ideas for further elaboration, something which would significantly reduce the number of ideas. The article suggests an alternative design to crowdvoting, whereby the users assess the relevant criteria.

Introduction

A major challenge facing organizations is assessing new product/service ideas in an efficient and comprehensive manner (Pisano and Verganti, 2008). One approach to this is outsourcing, at least the initial idea screening, to potential customers and users, so-called crowdvoting (Toubia and Florès, 2007). Empowering the crowd to assess new ideas could also have other positive effects such as strengthening a company's competitive advantage by means of their potential customers feeling involved and acknowledged (Fuchs and Schreier, 2011). However, there are also some potentially negative aspects of inviting the crowd to assess ideas. Companies can be expected to want to have final decisions made by in-house experts. The result of crowdvoting is thus seen, from a corporate perspective, as advisory, most definitely in the case of more complex and/or radical products (Pisano and Verganti, 2008). However, there is a risk of the crowd expecting the hosting company to implement the promoted ideas. Ideas often lack business potential, especially those emanating from outside the company (Bayus, 2013). If a company ignores the crowd's decision, it will risk causing bad will (Di Gangi, Wasko, and Hooker, 2010). Accordingly, for companies conducting or planning crowdvoting, it would be valuable to establish conformance between the evaluations of professional in-house experts and users. Despite a comprehensive review, no studies were found which were aimed at investigating conformance between user assessments and the assessments of professional in-house experts. Some related research was found, however.

In a field experiment, Onarheim and Christensen (2012) established conformance between a group of executives screening selection and an individual screening conducted by a group of employees, whereby more experienced employees conformed better. Toubia and Florès (2007) simulated different algorithms for adaptive idea screening. Their results indicate that a large number of consumers in idea screening could be compared with involving a small number of experts when making binary holistic evaluations, i.e., an idea could be either

“good” or not. Finally, in an experiment featuring MBA students as user evaluators, Ozer (2009) found that both *product expertise* (supply-side knowledge), and *product lead-usership* (demand- and supply-side knowledge) were related to more accurately predicting product success. When comparing both, product lead-usership was the more important. No comparison was made with expert evaluations, however.

Although contributing valuable knowledge to idea screening, none of those studies has thoroughly compared the users’ assessments with the true experts’ assessments. The purpose of this article is to investigate the appropriateness of outsourcing the initial idea screening of new product/service ideas to users as a means of relieving the workload of professional experts. In other words, the research question can be formulated thus: Are users’ assessments during idea screening conformant with professional experts’ assessments and are they reliable as a proxy for experts during idea screening?

This article reports on a comparative study where the idea screening of wireless services in mobile telephony was investigated. Screening was outsourced to two different types of users, (1) Technically-Skilled Users and (2) Technically-Naïve Users, and compared with the assessment made by professional experts. The rationale for this grouping of the users was theoretically founded. The main difference between the two groups was the degree of technology knowledge (supply-side knowledge).

The article starts by presenting the theoretical background; theoretical foundations for evaluating ideas, elaborating on the characteristics of judges that could theoretically influence assessments, and more specifically on the expected appropriateness of users. The theoretical framework concludes with a number of hypotheses regarding the hypothesized assessment conformance between users and experts.

Based on the hypotheses, a matching research design is described, subsequently followed by the results of the statistical analyses. These are then discussed, along with the managerial

implications linked, in particular, to the issue of crowdsourcing and crowdvoting, the limitations, and the implications for future research.

Theoretical Background

First the concept of crowdvoting is reviewed and defined. The characteristics relevant to a good judge are then discussed. These are then applied to users in order to discuss their appropriateness as judges. Two different types of users (Technically-Skilled, and Technically-Naïve) are established from theory. Finally, based on the review, a number of testable hypotheses are generated concerning how users would be expected to assess ideas in comparison with professional experts.

Crowdvoting

Crowdsourcing (Howe, 2006) has been gaining increasing interest due to involving customers and users in the contribution of novel ideas. Examples of this phenomenon include Dell's "IdeaStorm," Nokia's "Ideas Project," Sara Lee's "Making Connections," and "Threadless" where everyone is encouraged to contribute ideas for future improvements, new features, products, and services. Successful sourcing can result in such a large number of ideas that screening and selecting the best ones becomes difficult and costly (Di Gangi, Wasko, and Hooker, 2010; Pisano and Verganti, 2008; Soukhoroukova, Spann, and Skiera, 2012). Accordingly, 'crowdvoting' has also been gaining increasing interest as regards assessing new ideas. Crowd voting has a kinship with the concept of the 'Wisdom of the crowd' (WotC), coined by Surowiecki (2004). The underlying rationale being that independent judgments originating from a crowd are relatively accurate. According to Pisano and Verganti (2008), the crowd is better suited to deciding simpler aesthetic issues, e.g. the preferred design of a selection of objects, rather than choosing between solutions to technical problems. An example of the former is Threadless.com, which utilizes users in contributing

new T-shirt designs, in addition to selecting the best ones for mass production (Piller and Walcher, 2006). In this case, it is quite obvious that the crowd's selections will provide the most likely market success as the people voting are also the potential buyers. When it comes to more complex issues, single popular voting is often insufficient as regards capturing complexity (Pisano and Verganti, 2008; Riedl et al., 2013).

As mentioned in the introduction, there are risks attached to crowdvoting, mainly in that the crowd will expect their choice to also be adopted and realized by the hosting company. If this is not the case, discontent on the part of the crowd can result (Di Gangi, Wasko, and Hooker, 2010). Previous research indicates that the crowd and professional experts have different frames of reference when evaluating how good an idea is. For instance, the crowd does not have to consider issues such as feasibility, or business potential, when assessing an idea (Bayus, 2013).

Despite the popularity of actually engaging users in crowdvoting, there has been little investigation of users' appropriateness when it comes to judging the submitted ideas. As the literature review shows, remarkably little attention has been paid to investigating the characteristics deemed appropriate in people judging new product/service ideas. There is, thus, a need to better understand if and why the relevant dimensions are being assessed similarly, or differently, by users and professionals. To better understand this issue, we first investigate which characteristics are relevant to a judge.

Which characteristics are relevant in a judge?

In order to develop a theoretical understanding of the appropriate idea assessment characteristics of judges, the abilities necessary to understand and evaluate innovation are analyzed in detail.

The literature makes the basic assumption that a judge who is appropriate for idea screening is also an expert in the field (Amabile, 1996; Hogarth, 2001; Ozer, 2009; Salas,

Rosen, and DiazGranados, 2010; Shanteau, 1992). Research has also shown that expertise, competence, and experience are positively related to better decision-making as regards future predictions (Chi, Glaser, and Farr, 1988; Hogarth, 2001; Lovett and Anderson, 1996; Salas, Rosen, and DiazGranados, 2010; Shanteau, 1992; Terwiesch and Xu, 2008), even if the degree of novelty makes the prediction harder (Onarheim and Christensen, 2012). Various types of expertise have also been discovered to be important; product expertise (supply-side knowledge) and product lead-usership (demand- and supply-side knowledge) are related to more accurate predictions, with the latter being the more important (Ozer, 2009).

The importance of balancing supply- and demand-side knowledge has been acknowledged as important to success in the innovation literature (Lüthje, 2004; Magnusson, 2009; Reid and de Brentani, 2004). Ultimately, it comes back to the nature of new product/service ideas. These can be seen as consisting of two entities, i.e., a problem and a solution (Toubia and Florès, 2007; von Hippel, 1994), with these two parts corresponding to the supply- and demand-sides. Supply-side knowledge – also called technical knowledge (Lüthje, 2004), technology domain knowledge (Reid and de Brentani, 2004), and technology knowledge (Magnusson, 2009) – is necessary in order to understand how to implement ideas (the solution) and consists of the ability to analyze technical feasibility, or the opportunities and limitations of a given technology. Following Perrow (1967), we are of the opinion that ‘technology’ should be understood in terms of including product-specific technology, service-supporting technology, and organizational routines, thus including all the organizational resources necessary to produce products and services, i.e., all the enabling resources.

Demand-side knowledge – also called use experience (Lüthje, 2004), application domain knowledge (Reid and de Brentani, 2004), and use knowledge (Magnusson, 2009) – involves understanding how user/customer benefits (the problem) are created when using the underlying technology; it consists of knowledge relating to users’ needs and wants, together

with an understanding of how the product or service creates value for the user (Lüthje, 2004; Vargo and Lusch, 2004; Woodruff and Flint, 2006).

To conclude, the supply- and demand-sides are complementary types of knowledge which are necessary in order to accomplish and, accordingly, evaluate innovation (von Hippel, 1994). In light of this, the appropriateness of users as evaluators will now be discussed.

Appropriateness of users as evaluators

Users cannot be seen as a homogenous group when acting as assessors. Drawing on the works of Rogers (1995), concerning the diffusion of innovation, users can be divided into five different categories depending on their adoption behaviour. The two first adopter categories, *innovators* and *early adopters*, have also been described as being on the left side of the 'chasm' (Moore, 2005; Trott, 2012). The chasm is defined as the demarcation between two fundamentally different types of users, i.e., those to the left (*innovators* and *early adopters*) wanting technology and performance, and those to the right (*early majority*, *late majority*, and *laggards*), where potential customers are more interested in solutions and convenience, e.g., what the technology can do for them. The division into types of users has a kinship with the 'Bass model,' which defines users as being either innovators or imitators (Bass, 1969).

The innovators and early adopters are thus leading-edge users who are interested in new technology and constantly on the look-out for new and improved performance, even creating their own innovations to satisfy these needs. They possess greater technology knowledge than the other categories. For a long time, the innovation literature has recognized their potential ability as innovators, known as 'lead users' (Franke and Shah, 2003; Lüthje, Herstatt, and von Hippel, 2005; Urban and Von Hippel, 1988). It should be noted that only a fraction of the leading-edge users also are lead users. Lead users' abilities as evaluators have been sparsely described in the literature. However, Pitta and Fowler (2005) contended that online consumer

communities, and especially lead users, can contribute positively to the screening of new ideas as they are constantly engaged in discussing both new technology and its application within the scope of the community. Furthermore, when engaged in crowdsourcing, users tend to gradually learn and adapt their ideas so that these become more feasible (Huang, Singh, and Srinivasan, 2011). Thus, more experienced users can be expected both to learn and to gain more technology knowledge.

The three remaining categories are on the right-hand side of the chasm, being described as less interested in technology and thus expected to possess less technology knowledge. They have been described in terms of not having sufficient technology skills to come up with new, realizable solutions (Bennett and Cooper, 1981; Christensen and Bower, 1996). Accordingly, these technically-naïve users are mostly interested in the use-side of technology (Moore, 2005; Trott, 2012). A competent assessor must have enough use knowledge to understand the problem, focusing most interest on the early majority users of the three right-hand chasm categories. However, the technically-naïve users have also been acknowledged as having the potential to create new product/service ideas (Kristensson, Gustafsson, and Archer, 2004), even though Magnusson (2009) found their solutions often too immature and incomplete for direct implementation. Furthermore, he also found that company professional experts tended to emphasize feasibility issues at the expense of user needs when creating ideas, i.e., the tendency to create implementable but not so user-valuable ideas. In contrast, technically-naïve users tend to be better at understanding the demand-side than the supply-side.

Based on the review, users can be divided into two groups on the basis of both their use knowledge and their technology knowledge: '*Technically-Skilled Users*' and '*Technically-Naïve Users*'. The main difference between these is that the Technically-Skilled Users have a higher degree of technology knowledge than the Technically-Naïve ones. Their needs could

also be expected to be somewhat different, according to diffusion theory (Rogers, 1995), as previously explained.

All in all, this has implications for the expected ability to assess the different dimensions of product/service ideas that can be formulated using a number of hypotheses regarding the conformance between user judges and professional experts.

Hypotheses

In order to compare the conformance of the users with the professional experts, adequate measures must first be established which reflect dimensions relevant to idea screening. Considering the complexity of the task of evaluating a heterogeneous set of ideas, such as in our case involving wireless services, the literature recommends that it be broken down into the evaluation of separate dimensions, or criteria, instead of trying to make holistic decisions (Kleinmuntz, Fennema, and Peecher, 1996; MacGroger, Lichtenstein, and Slovic, 1988; Ozer, 2008; Riedl et al., 2013).

Numerous studies have looked into which criteria to use when evaluating new product/service ideas (Baker and Albaum, 1986; Balachandra and Friar, 1997; Carbonell-Foulquie, Munuera-Aleman, and Rodriguez-Escudero, 2004; Hart et al., 2003; Hauser and Zettelmeyer, 1997). To summarize this stream of research, there are no uniformly-accepted criteria for idea screening; criteria should be chosen with regard to the given context and phase of the development cycle (Hart, 1993). Drawing on previous experimental research studies that have investigated idea screening (Kristensson, Gustafsson, and Archer, 2004; Magnusson, 2009; Poetz and Schreier, 2012; Riedl et al., 2013), three criteria dimensions can be considered during the idea screening phase as regards assessing the merits of product/service ideas. These criteria dimensions are: originality, user-value, and producibility.

All together, the three dimensions listed above represent important aspects of successful innovation; i.e., something that is novel, useful, and possible to realize.

Originality, representing novelty and innovativeness; that is, how unusual, unique, and fresh the ideas regarding the prospective service can be considered. Drawing on Amabile (1996), we agree that originality is a concept which is generic and that people have an intuitive feeling for what is creative; it can thus be expected that this dimension is conformably assessed by all types of judges.

User Value, representing the users' perspectives regarding whether or not the implemented service ideas will create value for them. User Value corresponds to the demand-side. It can be expected that the Technically-Skilled Users (being on the left-hand side of the chasm) have more in common with the expert judges as they might tend to include more leading-edge values than the Naïve Users.

Producibility, representing the producer's perspective regarding the ease with which the service can be implemented and produced; that is, whether or not realizing the idea and charging for it will be possible. This dimension takes the supply-side perspective. To assess this, technology knowledge is required, something that is missing among the Naïve Users. However, the Technically-Skilled Users could be expected to have sufficient technology knowledge to conform with the Experts.

From the above, it can be hypothesized that the Technically-Skilled Users should be conformant with the Expert Judges in all three dimensions when ranking the relative merits of a set of ideas.

(H1) The relative scores assigned by Technically-Skilled Users will be conformant (co-vary) with the relative scores assigned by Expert Judges in regard to Originality, User Value and Producibility

The Technically-Naïve Users could, however, be expected to assess things differently. From the review, it can be predicted that the Technically-Naïve Users have other perceptions as regards what creates value in a product/service than the professional experts. Furthermore, their low level of technology knowledge will most likely cause them problems when assessing producibility, in the same manner as a professional expert. However, as previously stated, originality can be expected to be assessed as conformant. It can thus be hypothesized that:

(H2) The relative scores assigned by Technically-Naïve Users will be conformant (co-vary) with the relative scores assigned by Expert Judges in regard to Originality

(H3) The relative scores assigned by Technically-Naïve Users will not be conformant (co-vary) with the scores assigned by the Expert Judges in regard to User Value and Producibility

So far, the relative level of conformance between the users and the expert judges have been considered. Relative comparison compares the rankings of a set of selected ideas. When comparing absolute means instead the judges ratings being directly compared on an absolute level. Absolute scores are affected by the individual judge's previous experience. It is unlikely, for instance, that a Technically-Naïve User would assign the same absolute scores on originality as a professional expert. The experts have much more experience from various new products at the forefront, and are thus more conservative in their judgements. It is likely that both the Technically-Naïve and the Technically-Skilled Users will assign other absolute scores to individual ideas than the professional experts. Regarding absolute conformance, it can thus be hypothesized that:

(H4) Technically- Naïve Users will assign significantly different absolute scores than the Expert Judges in regard to Originality, User Value and Producibility

(H5) Technically-Skilled Users will assign significantly different absolute scores than the Expert Judges in regard to Originality, User Value and Producibility

Furthermore, in a real situation, it will mostly be of interest if the users are able to select the same “top ideas” as the experts. The ability applicable to such a setting is harder to predict. Instead of hypothesizing, we chose to explore this. With this in mind, a research design was formulated and is now described.

Study Method

An experimental design was used which employed real experts and users to evaluate wireless services. The experts were all experienced professionals within the field of wireless communications. The users were recruited to represent either Technically-Naïve or Technically-Skilled Users. All participants assessed a set of ideas in terms of three formal criteria (originality, user value, producibility). Their ratings were then compared in terms of absolute, relative, and top-selection conformance.

Idea collection

The ideas used in the study were collected from a commercial web portal operated by a telecommunications company, Telia, which invited mobile phone users to suggest ideas for future mobile phone services. Telia is a Swedish-based international telecom operator founded in the 1850s, which today is a part of TeliaSonera with more than 25,000 employees and net sales of approximately USD 14 billion in 2013. It is a world-leading pioneer in wireless services, being at the forefront of developing and operating wireless networks. Particularly in the wireless area, the company has been considered to be at the forefront of

innovation. The company launched one of the first commercial, automatic mobile telephony systems in the world in 1951. Its R&D department has been one of the driving forces behind the development of many successful wireless technologies and standards, e.g. NMT and GSM.

The ideas were submitted between mid-2008 and the beginning of 2009. It is noteworthy that this period occurred before the breakthrough of smartphone apps. The web portal was open to the public and allowed anyone to create an account, present ideas, and comment on other users' ideas; however, the ideas could not be rated in terms of figures or kudos. Users of the web portal were asked to share their ideas for future services regardless of their state of development, resulting in a level of detail ranging from fragments of ideas, based on current needs, to exhaustive presentations of possible services. This study includes a subset (approximately 25%) of the ideas from the web portal which was based on the following exclusion and inclusion criteria. Firstly, ideas that were advertisements by third-party developers, rather than novel ideas, were excluded. Secondly, the remaining idea providers were contacted and asked to consent to their ideas being used in the research. Only the ideas by idea providers responding positively to the invitation were included in the final sample, which consisted of 83 ideas contributed by 47 unique idea providers (37 male and 10 female, mean age = 38.9, SD = 10.5).

Participants and panels

The following three panels, representing different approaches to idea evaluation, were used: (1) Professional experts, (2) Technically-Skilled Users, and (3) Technically-Naïve Users. These panels are briefly described below.

Professional Experts (3 male and 1 female, mean age = 37.8, SD = 3.9), employed by the web portal provider, in the field of assessing and developing wireless services, were recruited in order to judge the ideas. All four had more than five years' experience of evaluating

wireless ideas. These four experts actually constituted the entire staff employed in evaluating incoming wireless ideas. One of the judges managed and maintained the web portal while the other three were employed on other projects. The company's strong innovation tradition, together with the working experience of the professionals, ensures that these judges can be considered qualified experts within the field.

Technically-Skilled Users (19 male, mean age = 28.7, SD = 10.8) were voluntarily recruited from an Internet-based community of interest focusing on Android mobile phones. The Android system was launched in Europe in early 2009, with the user community starting up in mid-2009. Today, the community has over 150,000 unique visitors each week. Furthermore, the community also has over 47,000 active members who have generated more than 1.4 million posts. Since volunteer recruitment was carried out in early 2010, it can be argued that the user community should be seen as a 'den' of pioneering users and early adopters, and thus consisting of people with a high level of interest in new technology, i.e., Technically-Skilled Users.

Technically-Naïve Users (4 male and 7 female, mean age = 23.7, SD = 2.4) were voluntarily recruited to represent non-technically-skilled users. None of these users had any in-depth knowledge of how wireless communication actually works. Nevertheless, they had all been using mobile phones in their day-to-day lives for several years.

Idea assessment

All ratings were done using a web-based tool through which the included ideas were presented to the judges, one at a time and in random order. Note that no comments from the web-portal community were included or displayed to the judges. The judges entered their responses, individually and independently, using a sliding-bar-scale from 1-100, at their own convenience. Once they had assessed all the ideas, they were asked to complete a short

demographic questionnaire which also included questions on technological knowledge and use knowledge vis-à-vis mobile phones and mobile technology.

Measures

Three different criteria were used when assessing the ideas: Originality, User Value, and Producibility. The criteria and the rationale behind using these measures were explained in the theoretical background. Each idea was rated on a scale from 1 to 100, with 1 being the lowest and 100 being the highest, on a visual analogue scale (Butler, 1997). For the data analyses, the data from each panel was averaged so that each panel received one composite score for each idea for each of the respective measures.

Manipulation check

In order to investigate whether or not the panels were indeed different, they were compared in terms of use and technological knowledge. A one-way between-groups ANOVA, with the three panels [Expert User, Technically-Skilled User, Technically-Naïve User] as the independent variable and technological knowledge as the dependent variable, showed a significant $F(2,31)24.852$, $p<.001$ main effect. A follow-up Tukeys HSD *post hoc* test showed that the Expert Users ($M=5$, $SD=0.0$) reported a significantly ($p<.001$) higher level of technological knowledge than did the Technically-Skilled Users ($M=3.21$, $SD=0.63$), who in turn reported a significantly ($p=.008$) higher level of technological knowledge than did the Technically-Naïve Users ($M=2.45$, $SD=0.69$). Furthermore, a one-way between-groups ANOVA, with the three panels [Expert Users, Technically-Skilled Users, Technically-Naïve Users] as the independent variable and use knowledge as the dependent variable, showed a significant $F(2,31)6.384$, $p=.005$ main effect. A follow-up Tukeys HSD *post hoc* test showed that the Technically-Skilled Users ($M=4.95$, $SD=0.12$) reported a significantly ($p=.004$)

higher level of technological knowledge than did the Technically-Naïve Users ($M=4.39$, $SD=0.71$), with the Expert Users placing between the two ($M=4.92$, $SD=0.17$).

Inter-judge reliability

To assess whether or not the judges understood the instructions in the same way, and subsequently evaluated the ideas in a coherent fashion, the three scales Originality, User Value, and Producibility were evaluated using Cronbach's Alpha (α_c) and Krippendorff's Alpha (α_k). The main difference between the two measures is that, while α_k is based on disagreement between scores, α_c calculates consistency without taking into account the actual levels of the scores. In other words, α_c can be said to evaluate whether or not the judges have understood the scale's underlying construct and subsequently assessed the ideas in a consistent fashion, while α_k additionally evaluates whether or not the judges share the same end-points of the scale. Thus, α_c is a good measure of the reliability of aggregated measures based on many observers, while α_k is better if the objective is a measure where absolute agreement is of interest (Hayes and Krippendorff, 2007). As no attempt was made to calibrate the judges' interpretations of the end-points of the scale, as advocated by, for instance, Amabile (1996), both the absolute α_k and relative α_c measure of consistency are presented below.

For the three criteria, the α_c values ranged from .743 to .786 for the Professional Experts (see Table 1 for values), from .686 to .861 for the Technically-Skilled Users (see Table 2 for values), and from .633 to .672 for the Technically-Naïve Users (see Table 3 for values). Note that deleting the least homogenous judge only had a minimal, positive effect on three of the α_c values of the scale, and even had a negative effect on two of the scales. According to the rules of thumb suggested by George and Mallery (2003), values $> .9$ are excellent, $> .8$ good, $> .7$ acceptable, $> .6$ questionable, $> .5$ poor, and $< .5$ unacceptable. Thus, the variation in responses does not seem to stem from any specific judge but from the complexity of the tasks

and the differences in frames of reference amongst the three panels. This interpretation is corroborated by α_k values which ranged from .289 to .460 for the Professional Experts (see Table 1 for values), from .060 to .223 for the Technically-Skilled Users (see Table 2 for values), and from .095 to .127 for the Technically-Naïve Users (see Table 3 for values). According to Landis and Koch (1977), values $> .81$ indicate almost perfect agreement, $> .61$ substantial, $> .41$ moderate, $> .21$ fair, $> .01$ slight, and $< .1$ poor agreement.

Insert Table 1 here

Insert Table 2 here

Insert Table 3 here

Results

Relative conformance between panels (H1-H3)

A series of Pearson correlation analyses was used to investigate the extent to which the evaluations of the two user panels (Technically-Skilled Users and Technically-Naïve Users) corresponded to the scores of the Professional Expert panel. The correlation analysis between the Professional Experts and the Technically-Skilled Users showed the following results: Originality ($r = .544$, $p < .001$), User Value ($r = .803$, $p < .001$), and Producibility ($r = .378$, p

< .001) thus supporting hypothesis H1. The correlation analysis between the Professional Experts and the Technically-Naïve Users showed the following results: Originality ($r = .515$, $p < .001$), User Value ($r = .559$, $p < .001$), and Producibility ($r = .555$, $p < .001$), thus supporting hypothesis H2 but falsifying hypothesis H3.

In summary, the results showed that the two user panels (Technically-Skilled Users and Technically-Naïve Users) showed relative conformance with the Experts' evaluations.

Absolute conformance between panels (H4-H5)

To investigate the absolute conformance between the panels, that is, the extent to which the panels gave the ideas comparable scores, a series of repeated measure ANOVAs was employed using the Sidak-adjusted *post hoc* test.

Originality

A repeated measure ANOVA, with the three panels (Professional Experts, Technically-Skilled Users, and Technically-Naïve Users) as independent variables and Originality as the dependent variable, was run in order to investigate whether or not the panels had judged the ideas equally. The results showed a significant effect $F_{2,164} = (204.66, p < .001)$ between the panels. Follow-up pairwise comparisons (Sidak-adjusted) revealed that the highest scores were given by the Technically-Naïve panel, scores which were significantly higher than the Technically-Skilled panel's scores which, in turn, were significantly higher than the scores given by the Professional panel (all differences were significant at the $p < .001$ level). See Table 4 and Table 5 for means, standard deviations, and *Post hoc* test.

User Value

A repeated measure ANOVA, with the three panels (Professional Experts, Technically-Skilled Users, and Technically-Naïve Users) as independent variables and User Value as the

dependent variable, was run in order to investigate whether or not the panels had judged the ideas equally. The results showed a significant effect $F_{2,164} = (46.62, p < .001)$ between the panels. Follow-up pairwise comparisons (Sidak-adjusted) revealed that the highest scores were given by the Technically-Naïve panel, scores which were significantly higher than the Technically-Skilled panels scores which, in turn, were significantly higher than the scores given by the Professional panel (all differences were significant at the $p < .001$ level). See Table 4 and Table 5 for means, standard deviations and *Post hoc* test.

Producibility

A repeated measure ANOVA, with the three panels (Professional Experts, Technically-Skilled Users, and Technically-Naïve Users) as independent variables and Producibility as the dependent variable, was run in order to investigate whether or not the panels had judged the ideas equally. The results showed a significant effect $F_{2,164} = (12.97, p < .001)$ between the panels. Follow-up pairwise comparisons (Sidak-adjusted) revealed that the highest scores were given by the Technically-Naïve panel, scores which were significantly higher than both the Technically-Skilled and Professional panels (p 's $< .003$). See Table 4 and Table 5 for means, standard deviations and *Post hoc* test.

Insert Table 4 here

Insert Table 5 here

Taken together, the results of the three repeated measure ANOVAs support hypotheses H4 and H5, showing a highly-consistent pattern whereby having experience of technology leads to lower scores. This result is true for all pairwise comparisons on all three measures, except for the pairwise comparisons between the Technically-Skilled and the Professionals regarding Producibility, for which the scores were identical.

Top selection comparison

As previously discussed, in a real-life situation, a company will be most interested in selecting the “best” ideas and rejecting the not so promising ones. Therefore, this analysis focuses on the user panels’ ability to select the same best ideas as the experts. Some kind of weighting of the dimensions is necessary in order to compute the aggregated merit of an idea. Just how criteria weighting should be done is debatable and general guidelines are lacking (Soukhoroukova, Spann, and Skiera, 2012). In the innovation literature, innovations are often divided into two different types, incremental and radical (Crawford and Di Benedetto, 2000), both considered of value to an innovating company. Magnusson (2009) thus argued that different weightings ought to be done bearing in mind the type of innovation the company is aiming for. His line of reasoning was to propose two different “innovation indexes” for incremental and radical types of innovation; these give different weightings to the three criteria (originality, user value, and producibility) in order to create an incremental and a radical index for ranking the merits of ideas.

$$\text{Incremental_index} = .05*\text{originality} + .475*\text{user_value} + .475*\text{producibility}$$

$$\text{Radical_index} = .55*\text{originality} + .35*\text{user_value} + .10*\text{producibility} + 25.71$$

Accordingly, two top-selections were computed using the two different weightings of the three criteria.

An analysis was done which simulated the three groups selecting 50 percent of the best ideas for further development, and thus rejecting the remaining ones. Consequently, this was a test aimed at investigating the extent to which the panels select the same ideas for further investigation. The results showed that, with respect to the ideas selected on the basis of the Incremental index, a 57 percent overlap existed between the three panels. Furthermore, with respect to the ideas selected on the basis of the Radical index, a 50 percent overlap existed; see Figure 1. Consequently, in both cases, more than half of the selected ideas achieved consensus. As shown in the Venn diagrams (Figure 1), of the Professionals' top 50 selection, the users (Technically-Naïve + Technically-Skilled) selected 81 percent of the Radical ideas and 91 percent of the Incremental ideas.

Insert Figure 1 here

An in-depth analysis comparing the professional experts' ranking with the user groups' ranking showed that the Technically-Skilled Users selected the top 11 incremental ideas, and the top 9 incremental ideas, whereas the Technically-Naïve Users only managed to select the top 2 incremental, and top 3 radical. Table 6 and Table 7 show a comparison of the top 50% selection of the Professionals and the user groups for both incremental and radical.

Insert Table 6 here

Insert Table 7 here

The Technically-Skilled Users selected 79% (incremental), and 64% (radical) of the Professionals' top 50%. Correspondingly, the Technically-Naïve Users selected 67% (both incremental and radical) of the Professionals' top 50%. If both user-groups' selections were taken into account, 88% of the incremental and 81% of the radical ideas selected by the professionals would be selected. However, this would result in reducing the radical ideas by 31% and the incremental ideas by 35%.

Discussion and managerial implications

Despite the popularity of actually engaging users in crowdvoting, there has been little investigation of their appropriateness when it comes to judging submitted ideas. As the literature review shows, remarkably little attention has been paid to investigating the characteristics deemed appropriate for people judging new product/service ideas. The present study is unique as its contribution is new knowledge of investigating whether or not users can be used as a proxy for professional experts, i.e., if users can act as a substitute for in-house experts in making evaluations during idea screening. In doing so, this study deepens the understanding that users can be a substitute for experts as found by (Toubia and Florès, 2007). Their study used a rough estimate for deciding whether or not the idea would be good. However, this study actually looked into three established criteria used both in practice and in other research studies.

Furthermore, it contributes to the open and user innovation literature – which has mainly perceived users as the contributors of ideas or innovations. This study advances the notion that users can contribute to innovation by also participating in the assessment of ideas, thus decreasing workload. The findings herein constitute a foundation for further investigation of the subject. The main findings of, as well as the contributions made by, the article will now be discussed.

Users as proxies for expert judges

The analysis showed that none of the two user panels' (Technically-Naïve Users and Technically-Skilled Users) idea assessments, in respect of the three criteria, could be considered conformant with the professional expert panel in *absolute* scores, something which was also in line with our hypotheses (H4, and H5). In comparison with the experts, both user groups overestimated the ideas, thus giving them significantly higher scores than the experts. The most probable reason for this is that the users were less experienced in the field than the experts. In general, the experts are likely to have seen and evaluated more ideas, on average, than the average user, giving them a broader frame of reference. For example, an idea that appears original to a user might not seem as original to an experienced expert who has seen similar ideas. Consequently, this will result in a lower absolute rating by the experts.

However, the *relative* scores – how well the user panels were able to relatively rank the ideas in respect of the three criteria – showed that both user panels were conformant with the professional experts. In other words, although the absolute ratings from the user panels were not conformant with the experts' ratings, the relative ranking of the ideas could still be used to sort out the relatively best ones from a given number of ideas. This was also partly expected in the hypotheses. It was expected that the Technically-Skilled Users would be conformant in all three dimensions (H1) with the Professional Experts, as supported.

However, it was not expected that the Technically-Naïve Users would also be conformant with the experts regarding user-value and producibility, (thus rejecting hypothesis H3). One plausible explanation for this is that there is quite a difference between evaluating ideas and creating them. To create an idea with high user value, you will probably have to have the necessary use knowledge, i.e., an understanding of the users' needs, thus making users more apt than experts in respect of user value. However, even an experienced expert can probably understand whether or not an idea presented to them will bring value to its users. An

implication of this is that the experts might also, in fact, be able to assess user value; this despite that the literature often claims that, in order to understand the potential user value of an idea, users should be actively involved and inquire about the evaluation (e.g., Brady and Cronin, 2001).

The fact that the Technically-Naïve Users were also conformant (in relative measures) with the experts, when assessing the producibility of the ideas, is much harder to explain. It seems that even naïve users had some kind of comprehension regarding the difficulty of implementing the ideas they were evaluating; this despite not knowing enough to fully understand the underlying technology. We are, however, cautious as regards drawing any definite conclusions from this finding.

The study also had a more explorative part where the panels' criteria assessments were used to compute two indexes (incremental and radical). The intention was to emulate how the criteria assessments could have been used in a real setting to rank the ideas in respect of their merits for incremental, and radical, innovation. The results showed a fair level of conformance with the top choices made by the experts. The number of ideas could approximately be reduced by 50 percent, without risking the rejection of the experts' highest-ranked ideas. Employing a wider range of users (both Technically-Naïve and Technically-Skilled) could increase conformance but with less reduction of the ideas.

It should be noted that the top selection was an emulation using a pre-defined index. Anyhow, it provides an estimate of the conformance between the professional experts' selections and those made by the users. Further development of the weightings to be applied to the index could increase the level of conformance and should be an issue for future research.

Managerial implications - crowdvoting revitalized

In crowdvoting, users most often make binary assessments by either promoting or demoting ideas. This type of voting primarily reflects the popularity of an idea among its intended users. However, popularity might not be an adequate measure upon which to base an organization's decisions regarding idea selection (Bayus, 2013; Di Gangi, Wasko, and Hooker, 2010). For example, a company has to take into account strategic issues that may cause the rejection of even very popular ideas (see, for instance, Di Gangi, Wasko, and Hooker, 2010). Several scholars have also advocated the use of multidimensional scales for accomplishing better assessment quality when assessing more complex ideas (Ozer, 2008; Riedl et al., 2013). Furthermore, not all ideas match company strategy, even if they are perfect from an individual user perspective. Decisions based on strategic considerations might also be difficult to explain transparently because doing so may reveal a confidential strategy. Furthermore, empirical research has shown that rejecting ideas promoted during crowdvoting may create frustration among the crowd (Di Gangi and Wasko, 2009).

This article has investigated how to employ user assessment in a way that differs from the normal popularity voting process. Instead of letting users promote or demote an idea, this article used an evaluation model enabling users to evaluate ideas in relation to three relevant criteria. Criteria evaluation is conformant with how assessment is normally performed by organizations. Thus, this article did not investigate whether or not popularity voting by crowds is appropriate but whether or not users are appropriate substitutes for experts when judging ideas using three predefined criteria. The present study has indicated that users can act as a proxy for experts when using criteria assessment.

Basing crowdvoting on criteria assessment could be one alternative to popularity voting. One advantage of criteria assessment is that it does not reveal the popularity of individual ideas to the rest of the community, thus avoiding the problem of implementation expectations.

Users would still rate ideas, thus helping the company to explore the “best” ones. If the most popular ideas are not highlighted, it will also be more difficult for competitors to spot and steal the most popular ones, another problem being faced by innovating communities (Di Gangi, et al., 2010). However, one potential disadvantage is that users might be less committed to voting because the process is less transparent.

The restriction of user evaluation

The three evaluation criteria used in this study were derived from their usage in previous research and their aptness to be used by users; these are all intrinsic quality factors. Anyhow, strategic fit, which is a company-dependent factor (de Brentani, 2001), has been found to be an important criterion during initial screening to predict future product success (Carbonell, Escudero, and Aleman, 2004). This criterion would, however, be impossible for the users to evaluate due to them being unaware of the company's strategy. At some stage during the idea management process, this criterion needs to be applied; however, this task cannot be handed over to outsiders such as users. In an open idea management system, the strategic fit must be done after the users have made their assessments, thus risking users' verdicts having to be overruled by the company because an idea does not fit with the current strategy. To avoid this risk, a company could, before letting users assess them first, filter out ideas that do not have a strategic fit. It could, however, also be argued that strategic fit is a criterion that preserves the status quo at a company as it filters out ideas outside of the current strategy (March, 1991; Poetz and Schreier, 2012; Prahalad, 2004). If one instead seeks to break free from the current state, strategic fit as a criterion can indeed be a hindrance instead. We argue that strategic fit is not a measure of an idea's quality per se, just a measure of how well the idea fits with a company's current strategy. Strategic fit is thus a double-edged sword.

Limitations and Future Research

It is important to bear in mind that the discussion conducted in this article, regarding appropriateness, reflects how well user judges could replace professional judges during idea screening, as in the case of crowd judging. Thus, professional judges are assumed to be appropriate for assessing ideas. Of course, this assumption could be questioned, something which is, however, beyond the scope of this article. Nevertheless, because professional judges are normally used by organizations to assess ideas, the analysis is considered relevant. Furthermore, the expert judges used in the study were, in fact, true experts, as described in the text.

The study was designed using criteria evaluation, not popularity voting, which is quite common in many crowdvoting situations today. The recommendations are limited to the context of studying wireless services, which in any case is a huge market today. User-expert conformance was tested for three criteria (originality, user-value, and producibility). To investigate other criteria, new studies should be conducted. Conformance between users and experts was only established in the case of relative assessments, not absolute assessment. This entails that users can rank a set of ideas, and not assess a very limited number of ideas. Application would thus be about screening a huge number of ideas using many judges. This was also indicated by the good level of inter-judge reliability using Cronbach's Alpha, which tests for consistency without taking actual levels into account, and the rather poor Krippendorff's Alpha which takes the actual scores into account.

It could be argued that the difference between "absolute" and "relative" conformance is a matter of the level of aggregation. The huge span of the scale used (1-100) makes absolute conformance harder to accomplish. The inter-judge reliability would probably have been higher if a two or three level scale had been used instead since the absolute scores would then be filtered.

Future research should address the issue of whether or not professional experts are considered better judges than users. To accomplish such research, longitudinal empirical studies can be used to investigate the correlation between the idea screening results of different panels and the actual market success of ideas.

The generalizability of the findings to other contexts is an intriguing question. The discovery that even the naïve users conformed with the experts in assessing producibility is probably the most surprising finding. It might be that producibility is easier to gain an understanding of in the context of wireless services, as users over time uses the mobile phone and develop a feeling for what is possible, and what is not. One could thus expect the findings to be generalizable to other technology-based services. However, additional research is needed in order to reach conclusions concerning other contexts.

Conclusions

The results indicate that a company can employ users during the initial screening process using criteria assessment as a proxy for in-house experts in order to select the best ideas for further elaboration. This process would significantly reduce the number of ideas. This study showed that reducing incoming ideas by 50 percent would still retain the most promising ones for more thorough evaluation.

The study also indicated that users should be a mix of both technically-naïve and more technically-skilled because these two categories complement each other.

The literature shows that the flipside of popularity voting exists in that the crowd expects, or even demands, the most popular ideas to be implemented. However, this article suggests an alternative to popularity voting; namely, allowing users to assess important criteria in order to

avoid the popularity trap. However, this may provide less encouragement for user participation.

Acknowledgement

The research has been funded by grants from the Swedish Governmental Agency for Innovation Systems (VINNOVA) grant number 2006-02821, and the Swedish Knowledge Foundation (the SISB project). The authors wish to thank our colleague Dr. Lars E. Olsson for his helpful comments. We are also grateful for the comments and advices from the editor Gloria Barczak and the two anonymous reviewers.

References

- Amabile, T. M. 1996. *Creativity in context*. Boulder, Colorado: Westview Press.
- Baker, K. G., and G. S. Albaum. 1986. Modeling new product screening decisions. *Journal of Product Innovation Management*, 3 (1): 32-9.
- Balachandra, R., and J. H. Friar. 1997. Factors for success in r&d projects and new product innovation: A contextual framework. *IEEE Transactions on Engineering Management*, 44 (3): 276-87.
- Bass, F. M. 1969. A new product growth for model consumer durables. *Management Science*, 15 (5): 215-27.
- Bayus, B. L. 2013. Crowdsourcing new product ideas over time: An analysis of the dell ideastorm community. *Management Science*, 59 (1): 226-44.
- Bennett, R. C., and R. G. Cooper. 1981. The misuse of marketing: An american tragedy. *Business Horizons*, 24 (6): 51-61.
- Brady, M. K., and J. J. Cronin. 2001. Customer orientation: Effects on customer service perceptions and outcome behaviors. *Journal of Service Research*, 3 (3): 241-51.
- Butler, P. V. 1997. Linear analogue self-assessment and procrustean measurement: A critical review of visual analogue scaling in pain assessment. *Journal of Clinical Psychology in Medical Settings*, 4 (1): 111-29.
- Carbonell, P., A. I. R. Escudero, and J. L. M. Aleman. 2004. Technology newness and impact of go/no-go criteria on new product success. *Marketing Letters*, 15 (2-3): 81-97.
- Carbonell-Foulquie, P., J. L. Munuera-Aleman, and A. I. Rodriguez-Escudero. 2004. Criteria employed for go/no-go decisions when developing successful highly innovative products. *Industrial Marketing Management*, 33 (4): 307-16.
- Chi, M. T. H., R. Glaser, and M. J. Farr. 1988. *The nature of expertise*. Hillsdale, N.J.: L. Erlbaum Associates.

- Christensen, C. M., and J. Bower. 1996. Customer power, strategic investment and the failure of leading firms. *Strategic Management Journal*, 17: 197-218.
- Crawford, C. M., and A. Di Benedetto. 2000. *New products management* (6th ed.). Boston: Irwin/McGraw-Hill.
- de Brentani, U. 2001. Innovative versus incremental new business services: Different keys for achieving success. *Journal of Product Innovation Management*, 18 (3): 169-87.
- Di Gangi, P. M., and M. Wasko. 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): 303-12.
- Di Gangi, P. M., M. Wasko, and R. E. Hooker. 2010. Getting customers' ideas to work for you: Learning from dell how to succeed with online user innovation communities. *MIS Quarterly Executive*, 9 (4): 213-28.
- Franke, N., and S. Shah. 2003. How communities support innovative activities: An exploration of assistance and sharing among end-users. *Research Policy*, 32 (1): 157-78.
- Fuchs, C., and M. Schreier. 2011. Customer empowerment in new product development*. *Journal of Product Innovation Management*, 28 (1): 17-32.
- George, D., and P. Mallery. 2003. *Spss for windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Boston: Allyn & Bacon.
- Hart, S. 1993. Dimensions of success in new product development: An exploratory investigation. *Journal of Marketing Management*, 9 (1): 23-41.
- Hart, S., E. J. Hultink, N. Tzokas, and H. R. Commandeur. 2003. Industrial companies' evaluation criteria in new product development gates. *Journal of Product Innovation Management*, 20 (1): 22-36.
- Hauser, J. R., and F. Zettelmeyer. 1997. Metrics to evaluate rd&e. *Research-Technology Management*, 40: 32-8.
- Hayes, A. F., and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1 (1): 77-89.
- Hogarth, R. M. 2001. *Educating intuition*. Chicago and London:: The University of Chicago Press.
- Howe, J. 2006. The rise of crowdsourcing. *Wired Magazine* (14.06): 1-5.
- Huang, Y., P. V. Singh, and K. Srinivasan. 2011. Crowdsourcing new product ideas under consumer learning. *SSRN eLibrary*.
- Kleinmuntz, D. N., M. G. Fennema, and M. E. Peecher. 1996. Conditioned assessment of subjective probabilities: Identifying the benefits of decomposition. *Organizational Behavior and Human Decision Processes*, 66 (1): 1-15.
- Kristensson, P., A. Gustafsson, and T. Archer. 2004. Harnessing the creative potential among users. *Journal of Product Innovation Management*, 21 (1): 4-14.
- Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33: 159-74.
- Lovett, M. C., and J. R. Anderson. 1996. History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology*, 31 (2): 168-217.
- Lüthje, C. 2004. Characteristics of innovating users in a consumer goods field: An empirical study of sport-related product consumers. *Technovation*, 24 (9): 683-95.
- Lüthje, C., C. Herstatt, and E. von Hippel. 2005. User-innovators and "local" information: The case of mountain biking. *Research Policy*, 34 (6): 951-65.

- Magnusson, P. R., E. Wästlund, and J. Netz. 2016. Exploring users' appropriateness as a proxy for experts when screening new product/service ideas. *Journal of Product Innovation Management*, 33 (1): 4-18.
- MacGroger, D., S. Lichtenstein, and P. Slovic. 1988. Structuring knowledge retrieval: An analysis of decomposed quantitative judgments *Organizational Behavior and Human Processes*, 42: 303-23.
- Magnusson, P. R. 2009. Exploring the contributions of involving ordinary users in ideation of technology-based services. *Journal of Product Innovation Management*, 26 (5): 578-93.
- March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science*, 2 (1, February 1991): 71-87.
- Moore, H. 2005. Break-through innovations and continuous improvement: Two different models of innovative processes in the public sector. *Public Money and Management*, 25: 43-50.
- Onarheim, B., and B. T. Christensen. 2012. Distributed idea screening in stage-gate development processes. *Journal of Engineering Design*, 23 (9): 660-73.
- Ozer, M. 2008. Improving the accuracy of expert predictions of the future success of new internet services. *European Journal of Operational Research*, 184 (3): 1085-99.
- Ozer, M. 2009. The roles of product lead-users and product experts in new product evaluation. *Research Policy*, 38 (8): 1340-9.
- Perrow, C. 1967. A framework for the comparative analysis of organizations. *American Sociological Review*, 32: 196-208.
- Piller, F. T., and D. Walcher. 2006. Toolkits for idea competitions: A novel method to integrate users in new product development. *R&D Management*, 36 (3): 307-18.
- Pisano, G. P., and R. Verganti. 2008. Which kind of collaboration is right for you? *Harvard Business Review*, 86 (12): 78-86.
- Pitta, D. A., and D. Fowler. 2005. Online consumer communities and their value to new product developers. *Journal of Product & Brand Management*, 14 (5): 283-91.
- Poetz, M. K., and M. Schreier. 2012. The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management*, 29 (2): 245-56.
- Prahalad, C. K. 2004. The blinders of dominant logic. *Long Range Planning*, 37 (2): 171-9.
- Reid, S. E., and U. de Brentani. 2004. The fuzzy front end of new product development for discontinuous innovations: A theoretical model. *Journal of Product Innovation Management*, 21 (3): 170-84.
- Riedl, C., I. Blohm, J. M. Leimeister, and H. Krcmar. 2013. The effect of rating scales on decision quality and user attitudes in online innovation communities. *International Journal of Electronic Commerce*, 17 (3): 7-36.
- Rogers, E. M. 1995. *Diffusion of innovations* (4 ed.). New York: Free Press.
- Salas, E., M. A. Rosen, and D. DiazGranados. 2010. Expertise-based intuition and decision making in organizations. *Journal of Management*, 36 (4): 941-73.
- Shanteau, J. 1992. The psychology of experts: An alternative view. In G. Wright, and F. Bolger (Eds.), *Expertise and decision support*. New York: Plenum Press.
- Soukhoroukova, A., M. Spann, and B. Skiera. 2012. Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets. *Journal of Product Innovation Management*, 29 (1): 100-12.
- Surowiecki, J. 2004. *The wisdom of crowds : Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (1st ed.). New York: Doubleday :.
- Terwiesch, C., and Y. Xu. 2008. Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54 (9): 1529-43.

- Magnusson, P. R., E. Wästlund, and J. Netz. 2016. Exploring users' appropriateness as a proxy for experts when screening new product/service ideas. *Journal of Product Innovation Management*, 33 (1): 4-18.
- Toubia, O., and L. Florès. 2007. Adaptive idea screening using consumers. *Marketing Science*, 26 (3): 342-60.
- Trott, P. 2012. *Innovation management and new product development* (5th ed.). Harlow, England ; New York: Financial Times/Prentice Hall.
- Urban, G. L., and E. Von Hippel. 1988. Lead user analyses for the development of new industrial products. *Management Science*, 34 (5): 569-82.
- Vargo, L. S., and F. R. Lusch. 2004. Evolving to a new dominant logic for marketing. *Journal of Marketing*, 68 (January): 1-17.
- von Hippel, E. 1994. "Sticky information" and the locus of problem solving: Implications for innovation. *Management Science*, 40 (4): 429-39.
- Woodruff, R. B., and D. J. Flint. 2006. Marketing's service-dominant logic and customer value. In R. Lusch, and S. Vargo (Eds.), *The service-dominant logic of marketing: Dialog, debate, and directions*. Armonk, NY: M.E. Sharpe.

Table 1: Cronbach's Alpha (α_c) for *Professional Experts* (n=4), α_c if the least homogenous judge is deleted, and Krippendorff's Alpha (α_k).

Scale	α_c	α_c if deleted	α_k
Originality	.783	.808	.431
User Value	.786	-	.460
Producibility	.743	-	.289

Table 2: Cronbach's Alpha (α_c) for *Technically-Skilled Users* (n=19), α_c if the least homogenous judge is deleted, and Krippendorff's Alpha (α_k).

Scale	α_c	α_c if deleted	α_k
Originality	.861	.862	.198
User Value	.877	.878	.223
Producibility	.686	.725	.060

Table 3: Cronbach's Alpha (α_c) for *Technically-Naïve Users* (n=11), α_c if the least homogenous judge is deleted, and Krippendorff's Alpha (α_k).

Scale	α_c	α_c if deleted	α_k
Originality	.672	.684	.116
User Value	.654	.677	.127
Producibility	.633	.658	.095

Table 4: Mean (M) and standard deviations (SD) for the three panels of judges (*Experts, Technically-Skilled Users, and Technically-Naïve Users*) for the three scales (Originality, User Value, and Producibility).

	Professionals		Technically-Skilled		Technically-Naïve	
	M	SD	M	SD	M	SD
Originality	25.16	17.61	47.24	13.07	55.21	12.03
User Value	53.24	17.00	58.82	13.51	65.00	11.79
Producibility	59.01	18.36	56.66	9.53	64.82	11.26

Table 5: *Post hoc* test (Sidak-adjusted) comparing the panels' assessments in absolute figures

<i>Dimension</i>	<i>Panel (I)</i>	<i>Panel (J)</i>	<i>Mean diff. (I-J)</i>	<i>Sign.</i>
Originality	Technically-Naïve	Technically-Skilled	7.969*	.000
	Technically-Naïve	Professionals	30.052*	.000
	Technically-Skilled	Professionals	22.083*	.000
User Value	Technically-Naïve	Technically-Skilled	6.177*	.000
	Technically-Naïve	Professionals	11.756*	.000
	Technically-Skilled	Professionals	5.579*	.000
Producibility	Technically-Naïve	Technically-Skilled	8.165*	.000
	Technically-Naïve	Professionals	5.815*	.003
	Technically-Skilled	Professionals	-2.350	.519

Table 6: Comparison of the top 50% selection of the Professionals and the user groups
Incremental index Magnusson(2009)

Idea #		Pro	TNU	TSU
108	Contact list	1	***	**
114	Who's calling?	2	***	***
23	I'm late - I need to record that ...	3		*
61	Improvements/Mobile phone	4	**	**
72	Mobile phones and computer	5	*	*
88	Automatic Backup!	6	***	*
67	Bar-code reader	7		**
58	When You are in phone answer with ...	8	***	***
74	Time limited duration silent mode	9	*	*
59	Call a group of people	10	***	*
78	No subject	11	*	*
111	Photo storage for mobile phones	12		
105	Built in key	13		**
69	Busy states!	14	**	*
53	Voicemail	15	***	**
54	Pay for travel using my mobile phone	16	*	*
77	It would be nice.	17	***	**
63	See my kids	18	*	
145	Health Monitor	19		*
116	TV - guide	20		*
45	Messages to a group	21	**	*
89	Lost your number, no worries!	22	**	
104	Anti-virus etc	23	**	
44	Social Communicator	24		
49	IP-telephony mode	25	***	***
65	Parents' best friend?	26	*	***
24	I want to use my phone...	27		***
30	A historical, cultural, tourism guide!	28	***	*
32	RideSharing	29		
33	Travelling	30	***	**
13	Split phone-related bills ...	31		*
81	Show my number to my friends and fam.	32	*	*
66	Mobile silencer	33	**	***
28	Location Indicator	34		
146	Mobile baby monitor	35	*	**
97	Queueing	36		**
1	drive too fast applications	37		
14	Paying by mobile phone	38	**	***
102	Digital TV - Mobile TV	39	**	***
7	Using the mobile for money transfer	40		*
96	School - quiz	41	*	
91	Loyalty programs at a mobile reach!	42	***	**
			28	33

TNU=Technically-Naïve Users; TSU=Technically-Skilled Users
5 of the experts were missed by the users

***: among the top 10 ideas

**: among the top 20 ideas

*: among the top 50% ideas

Table 7: Comparison of the top 50% selection of the Professionals and the user groups
Radical index Magnusson(2009)

Idea #		Pro	TNU	TSU
15	Approve your credit card shopping	1	**	***
82	Credit card ON/OFF	2	***	***
97	Queueing	3	*	***
120	VR Interface mobile phone	4		*
78	No subject	5		*
90	Credit Sharing!	6	*	*
32	RideSharing	7	***	**
14	Paying by mobile phone	8	**	*
96	School - quiz	9	***	**
136	Trumpet input	10		
121	Shake it into life	11	*	
4	Don't drink and dial!	12	**	
50	Virtual Home Village for Dispersed...	13		
87	Stay connected, keep track	14		
112	I want wine...	15	**	
139	phone with pockets?	16		
69	Busy states!	17	*	***
105	Built in key	18	***	***
151	Speech recognition and SMS	19	*	***
94	Use your friend's mobile device on...	20		***
26	Using the mobile phone as a car par...	21	*	***
66	Mobile silencer	22	***	***
145	Health Monitor	23	**	**
133	Phone projection keyboard	24	*	
7	Using the mobile for money transfer	25	*	*
44	Social Communicator	26		
116	TV - guide	27		
108	Contact list	28		
68	Smart Shopping	29	*	
77	It would be nice.	30	***	**
2	Silent SMS reminder or NFC	31	*	
155	Need for mobile spontaneous interpreting	32	*	*
74	Time limited duration silent mode	33	*	**
81	Show my number to my friends and fam.	34		**
12	Toll road payment by mobile	35		**
64	Personal finance app - better ...	36	***	*
92	TALK IS CHEAP!	37		*
130	Printer for text messages?	38	**	
61	Improvements/Mobile phone	39	*	*
23	I'm late - I need to record that...	40	**	***
25	Can you guess better than me?	41		
33	Travelling	42	***	**
			28	27

TNU=Technically-Naïve Users; TSU=Technically-Skilled Users

***: among the top 10 ideas

**: among the top 20 ideas

*: among the top 50% ideas

Figure 1: Distribution of the top 50% of ideas among the two user panels and the professional experts applying innovation indexes from Magnusson (2009).