Journal of Engineering Design
Vol. 23, No. 9, September 2012, 660–673

Taylor & Francis
Taylor & Francis Group

Distributed idea screening in stage–gate development processes

Balder Onarheim* and Bo T. Christensen

Department of Marketing, Copenhagen Business School, Copenhagen, Denmark

(*Received 24 February 2011; final version received 10 December 2011*)

This paper investigates the gate screening of ideas in engineering design, by examination of the validity of employee voting schemes and biases associated with such voting. After conducting an employee-driven innovation project at a major producer of disposable medical equipment, 99 ideas had to be screened for further development. Inspired by the concept of 'wisdom of the crowd', all ideas were individually rated by a broad selection of employees, and the ratings were used to investigate two biases in employee voting: visual complexity and endowment effect/ownership of ideas. The visual complexity bias was found to be a predictor for selection, but experienced employees were less affected by the bias. The ownership bias was potent in that every employee proved to be more likely to select his or her own ideas over other ideas, but this effect disappeared when aggregating across the crowd of employees. Furthermore, this study compared the employee selection with the preference of a small team of executives, showing that the employee voting significantly correlates with the preferences of the executives: overall, in the top 12 selected ideas and in the choice of idea categories. This match increases when including only the ratings of the most experienced employees.

Keywords: evaluation of creative ideas; creativity; idea evaluation; idea screening; engineering design; fuzzy front end; innovation; new product development; wisdom of the crowds; employee-driven innovation

1. Introduction

In complex and high-risk design projects, stage–gate models are often used as a management tool to control the development of the projects. At various gates, the design teams present a number of possible solutions, and a team of executive decision-makers discuss what ideas are to be developed further. The idea evaluation and selection, particularly in the early stages of engineering design, are critical, difficult and complex tasks (Cooper and Brentani 1984). Between the gates, the design teams use various iterative creative design processes to develop, evaluate and select different ideas and concepts to be formally screened in the gate meetings. Through this iterative process shifting between design work and gate selections, the number of ideas presented for screening at each gate meeting is reduced throughout a project, until the final concept is identified and realised.

While the generation of ideas in design processes has been examined in quite some researches, the evaluation process in formal idea screening (how are ideas to be evaluated, who is evaluating

*Corresponding author. Email: bo.marktg@cbs.dk

ISSN 0954-4828 print/ISSN 1466-1837 online

http://dx.doi.org/10.1080/09544828.2011.649426
http://www.tandfonline.com

and by which criteria) in early gates has not been the subject of much research, even though it represents an important element in the 'fuzzy front end' of new product development (Reid and de Brentani 2004). Only a pool of new ideas is an insufficient condition for innovation, as the importance lies equally in the recognition and selection of the best ideas (Rietzschel *et al.* 2010) to develop further. Having a number of good ideas generated does not matter if the right ones are not picked out for progression to later product development stages, and such selection is, therefore, a major challenge in the fuzzy front end of new product development (Soukhoroukova *et al.* 2012), particularly when, as is frequently the case, multiple potential ideas have been generated and company resources only permit very few of these to be developed further to become development projects. How are the right ideas selected for progression? Typical selection methods involve a selected few executives making the decision or a small panel basing their evaluation on inflexible criteria, such as what Cooper described as 'must have' and 'should have' (Cooper 2001, Cooper *et al.* 2002). Most theories of idea screening have focused on evaluations taking place at gates later in the innovation process, when initial ideas or projects have already been started (e.g. Cooper 2001). Stage–gate model theories tend to focus on the criteria to be applied in order to ensure that projects do not turn into runaway projects, in the sense that once started, there is a tendency to keep them alive and running much too long, at additional costs. Additionally, the portfolio management of the range of ideas that should enter into R&D projects has been examined (Cooper and Edgett 2007). An overview of previous research investigating methods for filtration and evaluation of new product ideas can be found in Crawford and Di Benedetto (2006).

In this paper, we challenge the approach of using only an executive team for idea screening in the early stages of product development, by comparing the selections of one such team with the ratings of a broad selection of employees, collected through individual voting schemes. This is in line with the claim made by Simmons *et al.* (2011): 'it seems obvious that companies should use the knowledge possessed by their employees during this fuzzy front end of new product development, but few organizations do so'. To date, much research has actually not examined the validity of such distributed but internal voting schemes for selecting ideas. To explore this validity, this study investigates the role of biases (e.g. idea ownership and expertise) in the selections made by employees. The use of distributed voting schemes for selecting ideas is inspired by theories of the 'Wisdom of the crowd' (WotC), stating that independent judgements of a crowd are relatively accurate under certain circumstances (to be reviewed below), even if many individuals in the crowd are error prone.

2. Literature review

2.1. *Idea screening*

In addition to the above-mentioned importance of idea screening in the fuzzy front end, idea selection, in general, is considered to be a notoriously difficult process. Studies have shown that not only do many companies lack a coherent or formal process for selecting ideas (Barczak *et al.* 2009), but people perform very poorly at selecting their own most promising ideas as well (Faure 2004, Rietzschel *et al.* 2006). Rietzschel *et al.* (2006) even found that the ideas selected for their creativity in some cases were no better than randomly sampling the pool of generated ideas! Finding clear criteria for selection has been pointed out by some scholars as an important step towards improvement of the quality of selected ideas (Cooper 2001, Rietzschel *et al.* 2010), but in a complex real-life context, finding the right criteria might be as challenging as selecting the best ideas. The lack of relevant and reliable data when screening product ideas (Cooper and de Brentani 1984) makes it challenging to know what criteria are to be focused on, and the consequences of choosing the wrong criteria can, of course, be fatal. Thus, it is no surprise that even a large

proportion of best practice companies acknowledge that they have problems with establishing clear criteria for product development processes (Cooper *et al.* 2002).

To add ecological validity to the research design, this paper utilises real-world engineering design problems and ideas from a large international company working in medical plastics. Is it possible to utilise a group of more or less randomly picked employees as an addition or supplement to other established evaluation methods?

2.2. *Distributed voting schemes*

In the last decade, there has been much focus on distributed innovation: inside companies, in between companies and from end users. These efforts have produced concepts such as crowdsourcing (Brabham 2008), idea sourcing (Joshi and Sharma 2004), idea markets (Soukhoroukova *et al.* 2012), co-creation (Prahalad and Ramaswamy 2004) and lead user innovation (von Hippel 1986). However, many companies operate in highly competitive markets where it is crucial to be able to patent innovations and to protect trade secrets, making it practically impossible to engage in major and critical innovation processes with external actors. Distributed innovation processes involving random samples or 'early adapters' from outside the organisation are, therefore, not always possible. In such cases, the companies have to focus on internal resources, and the company which served as the site for this paper (Coloplast) has recently started employing an employee-driven innovation (EDI) approach as a way of engaging in distributed innovation, albeit still ensuring that the involved participants are internal to the company (Onarheim 2011). At the core of EDI lies the principle of involving internal and already paid resources with broad and relevant domain knowledge, thereby excluding the potential risks associated with involving outsiders (e.g. legal ownership, intellectual property and trade secrets). Other important elements of EDI are motivation, social aspects, commitment and stakeholder management (Kesting and Ulhøi 2010, Onarheim 2011).

One possible element in an EDI process is the involvement of a broad selection of employees through distributed voting schemes. Such voting schemes, both inside and outside of an organisation, build on the idea of WotC (Simmons *et al.* 2011). The WotC hypothesis predicts that the independent judgement of a crowd of individuals (as measured by some form of central tendency) will be relatively accurate, even when most of the individuals in the crowd are ignorant or error prone (Surowiecki 2004, Simmons *et al.* 2011). For example, Galton (1907) famously reported that in a regional fair competition asking people to estimate the weight of an ox, the average estimate was just one pound short of the true weight of the ox. The WotC hypothesis implies that majority rule or average opinions will frequently outperform, as well as be more accurate in an absolute sense, decisions made by single judges and by experts or in-group decisions. The hypothesis is derived from mathematical principles, in that a crowd's judgement comprises signal plus noise, and averaging across judgements will then cancel out the noise while extracting the signal (Hogarth 1978, Simmons *et al.* 2011). The conditions for the occurrence of WotC are that (1) the crowd must be knowledgeable and (2) individual errors in judgement must not be systematic at the sample level. Systematic errors in judgement can, for example, occur with restricted diversity on the judging sample or lack of independence among the judges.

Compared with the typical setup of a WotC experiment, this study utilises the theoretical arguments from the WotC literature while applying it to a somewhat atypical sample. The pool of ideas that was used as data in this study was generated through an EDI process consisting of several workshops, and the employees used in the sample had all been part of this process. This increases the necessary level of knowledge in the sample, but also influences the level of diversity. Given the sample affiliation with the company in question, and the attendance of the employees in the workshops generating the ideas, it is likely that they are less diverse (at least

pertaining to domain knowledge) than a random sample, even though they were recruited from a wide range of departments, positions and levels of the organisation. Furthermore, the employee sample was smaller than what is typically used in most WotC studies. The WotC literature mainly focuses on randomly selected crowds in the number of thousands, so the 'crowd' in this study is both smaller and may be somewhat less diverse than a typical WotC sample. Still, for many, if not most, innovation- and technology-driven companies, it would not be possible to go much beyond this level of random evaluator selection due to the legal challenges mentioned above. Notwithstanding the differences , this paper argues that the advantages of potentially increased precision and added decision-making quality found in the WotC literature still stand in the sample used in this study. These theoretically driven notions of whether a particular sample is biased (e.g. in their kinds or degrees of diversity) in particular ways also drive this research: it is important to try to determine whether some form of bias may be leading the crowd to make erroneous or poor decision. This paper examines the potential impact on WotC by two sources of bias, as well as two ways to overcome them: visual complexity and endowment/ownership effects.

2.3. *Visual complexity in the information provided*

Some evidence from the creativity literature suggests that visual complexity may lead people to assume that the outcome is creative. Factor analysis has found that complexity loads on the same factor as originality and creativity (O'Quin and Besemer 1989, Young and Racey 2009), and recent research has shown how increasing complexity or lowering visual fluency leads to higher ratings of creativity (Christensen *et al.* under review) or product innovativeness (Cho and Schwarz 2006). As such, it is possible that individuals are using visual complexity as a heuristic for estimating product creativity and innovativeness – an important, and arguably the most important, criterion when estimating which ideas should be allowed to progress through gates in a product development process.

The use of such visual complexity heuristics for estimating product creativity or innovativeness may, however, be moderated by the level of experience of the judges. Experienced judges should be able to rely on more sources of knowledge of the market, of existing production methods, of the needs of the customer, of patented solutions and of competing and existing products on the market and should thus not have to rely on simply heuristics such as the link between visual complexity and creativity. Experts and novices often disagree systematically in their selections of product ideas (Moreau *et al.* 2001), and results from forecasting studies stress that using several experts instead of one leads to better results (Armstrong 2001). In addition, Cooper (2001) has argued for the need for experienced judges in evaluation in product development gates.

2.4. *Endowment/ownership effects*

It has been shown that when ideas are generated, the creators or contributors to the idea generation hold their own ideas in higher esteem than other ideas. In behavioural economics, this has been labelled the endowment effect, whereupon it has been shown that investing time and energy in developing a solution leads one to appreciate that solution more, and owning an object/solution leads to increased feelings of loss when having to let it go (e.g. Kahneman *et al.* 1991). Cooper (2001) described this as a problem in idea selection, in that it makes up a potential bias in screening ideas, thus prohibiting objective evaluations, and calls for the 'drowning of your puppies' in idea selection. Almost 50% of best practice companies in product development processes acknowledge that they have problems with the issue of establishing clear criteria and 'drowning their puppies' (Cooper *et al.* 2002). As such, the relation between who generated the ideas and who is to make the evaluation of which ideas should progress is important to consider. There

are different ways of trying to counter this well-known bias. Cooper *et al.* (2002) argued that it could be countered through the setup of clear selection criteria ('must have' and 'should have') to be implemented rigorously at the gates. However, such clear and rigorous criteria are hard to formulate unambiguously (which is why so many companies have a problem with implementing them), and furthermore, it is extremely difficult to find objective ways to weight these criteria against each other in the selection process (Soukhoroukova *et al.* 2012). An alternative way may be the use of WotC, by asking employees or other groups for holistic measures of overall promise for advancement in product development, for example, by voting or ranking the ideas. In an employee voting evaluation setup, the individual endowment effects should be cancelled out in the process, as long as there are no systematic endowment effects across the sample of raters. Systematic endowment effects across the sample of raters could, for example, occur if a large proportion of the raters were involved in the generation of a subset of the ideas, while others were generated by single individuals, or if the sample of raters represented a skewed proportion of the sample of generators (e.g. 11 groups of participants helped generate the ideas, but only 5 of these groups contributed to their evaluation).

3. Hypotheses

Based on the above literature review, this paper attempts to test the following hypotheses about biases in distributed voting schemes among employees:

H_1: High visual complexity in the presentation of an individual idea leads to more selections of that idea for further development.
H_2: Experienced raters should not rely on visual complexity heuristics to the same degree as inexperienced raters.
H_3: Ownership of an idea leads to a higher selection rate of that idea by individual raters.
H_4: The ownership bias should disappear while utilising WotC as long as the raters represent a random and unbiased sample of the subjects who generated the ideas.

In order to estimate whether these biases could be countered, the employees' ratings were compared against the choices of a team of senior marketers. Unlike other types of prediction markets, idea evaluation suffers from the fact that the ideas not chosen for progression cannot be evaluated *post hoc* (i.e. they drop out and are not developed further). Therefore, no objective measure exists to estimate the external validity of the selections of the crowd to what might have been selected (Kamp and Koen 2009). Previous researches estimating the validity of idea selection have utilised the same type of 'executive team' measure against which WotC could be measured and have generally found somewhat low levels of validity ranging 0.10–0.47 (LaComb *et al.* 2007, Soukhoroukova *et al.* 2012).

4. Methods

The design of this research was a real-world field study conducted in a major international producer of disposable medical equipment. The base for this study was a comprehensive 8-week EDI project at the company. In the project, 93 employees from 11 departments were involved in a total of 11 departmental and cross-departmental workshops, generating a pool of 99 distinct ideas described in writing and drawings and sorted in 26 different categories. As pointed out by Joshi and Sharma (2004), such a large number of contributors with diverse skills can enhance the chances of finding a truly innovative idea. It is important to note that all ideas were subject to ongoing and repetitive assessment and screening throughout the process leading to the final pool of ideas, an effort

that should ensure that all the 99 ideas should be considered of a satisfactory quality – not a random selection of 'all the ideas we could come up with'. Furthermore, the grouping of the ideas into categories was an emphasised part of the workshops, ensuring relevant and homogeneous categories. Categories of ideas were, therefore, used as a measure in this study, as it was assumed that ideas within one category have fundamental similarities. For a comprehensive description of the process of generation of the ideas, the executive selection and the rationale behind the layout, see Onarheim (2011).

4.1. *Materials*

Based on the output from the EDI process, the 99 unique product ideas generated were each described briefly in text, and the benefits to the company and to the end users were listed in a bullet-point fashion. Furthermore, the sketches were redrawn by a professional designer, resulting in a catalogue presenting all the 99 ideas in a standardised manner. Each idea was, as illustrated in Figure 1, presented on a horizontally oriented page, one half of the page with a short description of the idea in text, including a list of 'customer benefits' and 'company benefits', and the other half with drawing(s) and/or graphical figures. As all the ideas are potential future products for Coloplast, the ideas are considered confidential and the actual ideas are, therefore, unfortunately restricted from being shared.

4.2. *Measure of complexity*

Each idea was rated for visual, textual and benefit complexity by an independent researcher, unaware of the hypotheses of this paper. Visual complexity was counted as the number of separate

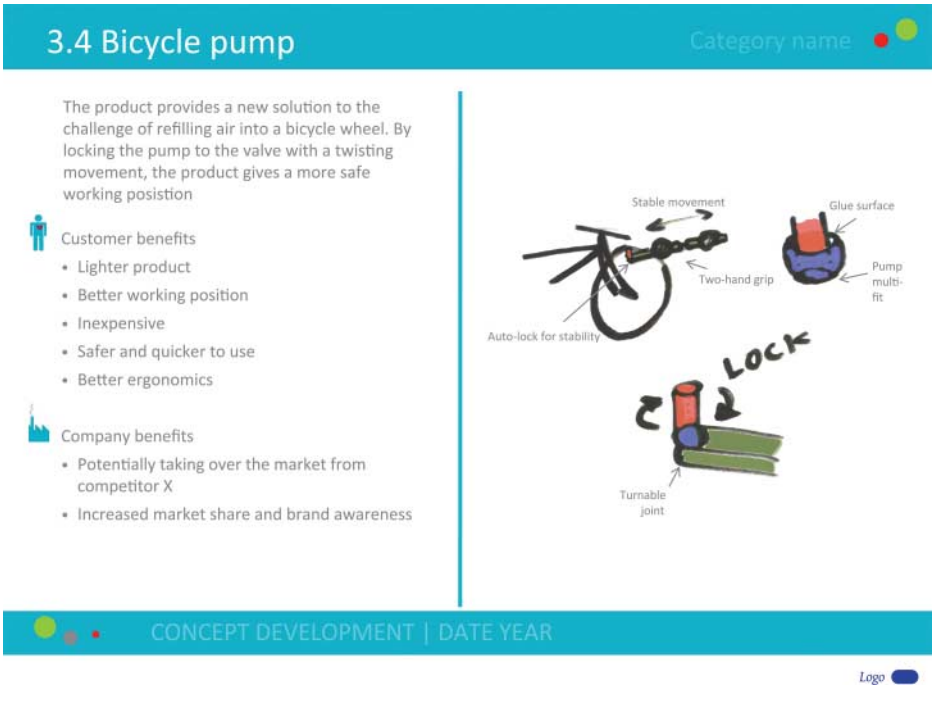

Figure 1. Template for the layout of the ideas evaluated by the employees.

drawings made to visualise the individual idea. Textual complexity measured the LIX value of the text describing the idea, calculated as $(O/P + L^*100/O)$, where O is the number of words in the text, P is the number of full stops and L is the number of long words, that is, with more than six letters. Benefit complexity was calculated simply as the number of bullet points describing the benefits of the idea for the company and the users.

4.3. *Product evaluation*

Two groups of company employees independently evaluated each idea. As part of the EDI process, company executives selected an executive team consisting of seven handpicked senior marketers representing four national subsidiaries and the global functions. The executive team thus consisted of a much less diverse sample than the employee sample, in terms of experience, seniority, placement in the level of the organisation, departmental affiliation, discipline training and background. Using such teams to select ideas for advancement is a usual way to filter ideas for new product development at the company. The executive team was gathered for a full-day workshop where all the ideas were assessed, with the criterion of finding the ideas best suited for further development. The team discussed what they considered important sub-criteria and reached a consensual understanding of what they considered important for the ideas selected. The workshop resulted in a selection of 12 ideas that were later turned into company product development projects. The executive team's evaluation served as the standard against which the wisdom of the employee crowd was compared to further investigate whether these two groups were in agreement. This is not based on an assumption that the executive team did a perfect job in their evaluation, but in this real-world project, it is the most accurate measurement as the ideas selected by the executives are the ideas that actually will be considered further. When ideas in product development are screened and some are discarded, there exists no objective knowledge about what would have been the best ideas (Kamp and Koen 2009), thus the selection of the executive team was used for comparisons in the employee voting scheme.

In addition to the executive team, the employees contributing ideas were invited to individually rate the 99 ideas through an online survey. Such a distributed assessment of ideas in the fuzzy front end is not a usual method employed at the company. The employee crowd evaluating the products was a group of 35 employees (16 females and 19 males, mean age 42) from a variety of job functions and company departments who had all taken part in the EDI workshops. They represented involvement from 11 departments and had a mean of 8 years of company experience (range 0–24) and 4 years of experience working in the product domain in question (range 0–24). The sample represented raters from all workshops and departments who had taken part in the EDI process. On the introduction page of the online survey, the participants were given the following instructions:

> On the next pages you will be asked to help evaluate the 99 individual ideas that came out of the workshops, based on the assumption that Coloplast does not have unlimited resources to develop all these ideas. Therefore, it is important to try to select the most promising ones that you think should be taken further in future development processes. As such, you should be critical in your selections, in order to ensure that the right ideas are selected for advancement. For each idea, please try to evaluate whether you think that Coloplast should develop and work on this idea (by answering yes/no). Also, for each idea you will be asked about whether you worked on the idea during the workshops (either by proposing it, or helping develop it further). If you worked on the idea, then please tick the appropriate box.

Under each presented idea, they answered the following: 'Would you recommend that Coloplast invest resources in order to try to develop this idea further?' [yes/no] and 'Did you work on this idea during the workshop?' [yes]. Each participant viewed all the 99 ideas one by one, randomised for ordering across participants, and answered the two questions. In order to analyse the voting behaviour of the employees, an average selection score was calculated for each idea. In order to compare these results with those of the executives, the top 12 ideas (corresponding to the number of

ideas selected for projects by the executives) were counted as the selection of the employee group. Furthermore, information about the level of expertise of the individual employees was obtained.

5. Results

5.1. *Descriptive statistics*

The mean number of times an idea was selected of the 35 raters was 14.6 (Standard Deviation (SD) 7.1, ranging from 1 to 32 and normally distributed). As such, no ideas were unanimously selected and all ideas were selected by at least one rater, confirming the assumption that all ideas presented can be considered somewhat relevant. The individual raters, on average, selected 41.3 ideas (of 99) for further work (SD 12.5, ranging from 15 to 67 ideas and normally distributed). This high number of ideas selected for continuation, and the fact that all ideas were selected at least once, indicates that all the 99 ideas presented were of a satisfactory quality.

5.2. *Inter-rater agreement*

The agreement among the raters was satisfactory. Intraclass correlation coefficient (ICC) (two-way for consistency) among the 35 raters for their selections was 0.87. It is possible to calculate how many evaluators would have been needed in order to reach a satisfactorily high level of agreement (ICC > 0.8) using a variant of the Spearman Browne Prophecy formula:

$$m = \frac{p^*(1 - p_L)}{p_L(1 - p^*)},$$

where m is the result to be rounded to the next highest integer, p^* is an aspiration level and p_L is a reliability estimate, typically either ICC(2,1) or ICC(3,1). For an experimental setup such as this (with 99 individual ideas to be rated and random judges), it is to be expected that to replicate the high level of agreement, 30 individual raters should have sufficed to reach a satisfactory agreement among the raters.

5.3. *Tests of hypotheses*

To investigate H_1, whether idea complexity biased subjects towards selection, the three kinds of complexity (textual, visual and benefit) were standardised and averaged across to generate a total complexity measure. A linear regression of whether the total complexity measure predicted the selection of the individual ideas producing an adjusted R^2 of 0.048 ($F(1, 98) = 5.98$, $p < .02$) with total complexity being a significant predictor ($\beta = 0.24$, $t(98) = 2.45$, $p < 0.02$) showed that idea complexity did predict selection.

To further examine H_2, whether employee expertise moderated the idea complexity bias, the employees were divided into two groups by expertise level with an approximate mean split. Expertise level was calculated by averaging the number of years of employment in the company and the number of years of experience in the product domain. The experienced group ($N = 15$) had a mean of 14 years of company experience and 8 years of domain experience, and the inexperienced group ($N = 20$) had a mean of 3 years of company experience and 2 years of domain experience. Experienced and inexperienced raters did not differ significantly in the mean amount of ideas they selected for further development from the set of 99 ideas (39 and 43 ideas selected for progression, respectively, $t(33) = 0.79$, NS).

For the inexperienced group, a linear regression of the three individual complexity measures (textual, visual and benefits) using a direct method showed an adjusted R^2 of 0.058 ($F(3, 89) = 2.89$, $p < 0.04$). Visual complexity ($\beta = 0.27$, $t(98) = 2.69$, $p < 0.01$) was significant, while textual complexity ($\beta = 0.12$, $t(98) = 1.20$) and benefit complexity ($\beta = 0.07$, $t(98) = 0.72$) were non-significant predictors. For the experienced group, a linear regression of the three individual complexity measures (textual, visual and benefits) using a direct method showed an adjusted R^2 of 0.076 ($F(3, 89) = 3.52, p < 0.02$). Benefit complexity ($\beta = 0.23, t(98) = 2.33$, $p < 0.03$) was significant, while visual complexity ($\beta = 0.19, t(98) = 1.88$) and textual complexity ($\beta = 0.16, t(98) = 1.54$) were non-significant predictors. In comparison, the same regression was run for the selection of the executive team, yielding an adjusted R^2 of 0.051 ($F(3, 89) = 2.64$, $p = 0.054$). Benefit complexity ($\beta = 0.23$, $t(98) = 2.24$, $p < 0.03$) was significant, while visual complexity ($\beta = -0.05$, $t(98) = -0.46$) and textual complexity ($\beta = 0.18$, $t(98) = 1.75$) were non-significant predictors. The results indicate that while both the executive team and the experienced group of employees relied slightly on the number of benefits indicated for each idea, the inexperienced group of employees did not consider the number of benefits in their selection, but instead relied slightly on visual complexity (the number of visual images shown).

To examine H_3, whether having worked on an idea biased evaluators towards selecting that idea, the proportion of ideas selected for the ideas the evaluator had or had not worked on, respectively, was calculated. Thirteen evaluators did not report having worked on any of the ideas, even though they had been present in at least one of the EDI workshop. A paired t-test showed a significant difference (paired $t(21) = 10.34$, $p < 0.001$), with a mean probability of picking an idea the evaluator had worked on of 0.81, with 0.40 for ideas not reported to have been worked on. The effect was so potent that for every evaluator, there were a higher average proportion of picks for ideas that had been worked on than ideas not worked on.

To estimate whether expertise mediated the ownership effect identified in H_3, an executive team was used as comparison. Overall the mean of the employee crowd selections correlated, $r(99) = 0.32$, $p = 0.001$, with the selection of the executive team. Besides the correlation, an important statistic in estimating validity is how many of the top picks (i.e. the ideas receiving the most votes) were actually shared between the executive team and the crowd. To estimate this, the 12 ideas (paralleling the 12 picks of the executive team) with the most votes were considered 'picks of the employee sample'. Furthermore, given that the ideas were sorted into 26 categories by overall topic, it was possible to also estimate how many of the general categories that the executives and the employee sample had agreed on selecting/not selecting ideas from. Among the top 12 executive picks, 5 of the ideas were also ranked in the top 12 picks by the employees (Cohen's $\kappa = 0.34$), and of the 26 categories of ideas in the pool, the two groups were in agreement in their picking/not picking an idea from a category in 21 of the categories (Cohen's $\kappa = 0.56$). Figure 2 shows the relation between the picks by the executive team and the top 12 picks by the employee voting scheme.

Figure 2 shows how employee ratings can be used to investigate executive selections. While the employees agreed with the executives on some ideas (e.g. 37, 64, 56 and 63), the employees preferred certain ideas not selected by the executives (e.g. 40 and 36) and disregarded some ideas selected by the executives (e.g. 99, 60, 81 and especially 26).

The same statistics were then computed for the experienced and inexperienced employees, respectively. Due to equality in the number of picks in some of the ideas in this reduced sample, the top 12 picks actually became the top 16 (experienced employees) and top 14 (inexperienced employees) in order to accommodate ideas with equal scores. In comparison, the experienced group of employees correlated, $r(99) = 0.33$, $p = 0.001$, with the executive team, shared 7 of the top 12 picks (Cohen's $\kappa = 0.42$) and agreed on picking/not picking 23 of the categories (Cohen's $\kappa = 0.75$), while the inexperienced group correlated, $r(99) = 0.29$, $p = 0.004$, shared 5 of the top 12 picks (Cohen's $\kappa = 0.29$) and agreed on picking/not picking 21 of the

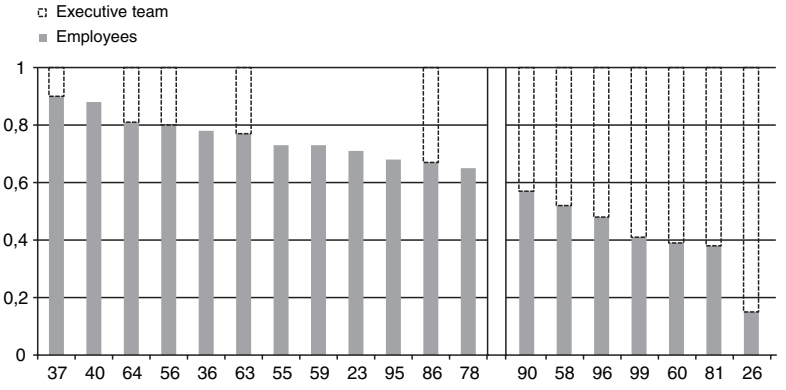

Figure 2. The executive team top 12 picks (white bars), the employee top 12 picks (left bars) and the respective proportion picks (grey bars) by the employees. The *x*-axis displays the unique idea number.

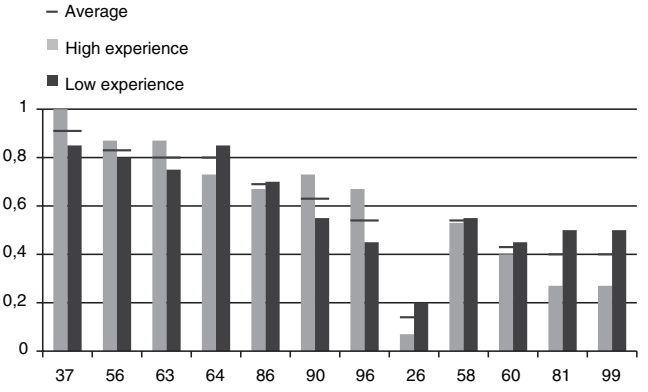

Figure 3. Proportion picks by experienced and inexperienced employees on the 12 ideas selected by the executive team for advancement.

categories (Cohen's $\kappa = 0.59$). The experienced and inexperienced employee groups correlated, $r(99) = 0.78$, $p = 0.001$, with respect to the selections, shared 11 of the picks in their top 14/16 (Cohen's $\kappa = 0.69$) and agreed on picking/not picking 20 of the categories (Cohen's $\kappa = 0.51$). As such, some significant gains appear to result from choosing experienced employees as evaluators – particularly in the shared picks in the top 12 picks and in choosing the general categories from which ideas would be selected in this case. To further illustrate the ratings of the experienced/inexperienced employees, the proportion of the two groups (high/low experience) that had selected the 12 ideas picked by the executive team was calculated, as shown in Figure 3.

This figure shows how half of the ideas picked by the executive team were selected by more than 60% of the employees. Five of the 12 ideas were preferred to a larger extent by the high-experience group (37, 56, 63, 90 and 96), while seven of the less popular ideas were preferred to a larger extent by less experienced raters (64, 86, 26, 58, 60, 81 and 99). Especially, ideas 26, 81 and 99 stand out as having a lower rating by the high-experience group, and these are also ideas with the lowest overall proportion picks. Idea 26 is the outlier, with only 14% picks in the average employee ratings, and 7% picks among the highly experienced raters.

To estimate H_4, whether the ownership bias would be countered in using an employee sample, all data on ideas the individual evaluator had worked on were excluded, and the above estimates of the correlation to the executive team were run once more. In effect, this did not change very

much in the results, probably due to the fact that most of the evaluators reported having worked on few ideas (six, on average), and these ideas varied from rater to rater. The correlation between the whole data set and the data excluding ideas that the individual reported to have worked on were $r(99) = 0.99$, $p = 0.001$, and the ranking among the top 12 ideas was almost identical, Spearman's $r_s(12) = 0.97$, $p = 0.001$. As such, the correlation of this new mean measure with that of the executive team did not show any differences compared with the correlation between the executive team and the full data set: the correlation was still $r(99) = 0.32$, $p = 0.001$ and there were still 5 shared picks in the top 12 (Cohen's $\kappa = 0.34$). As such, although the ownership bias was a potent one at the individual ratings, it was in support of H_4 cancelled out across the sample of raters.

5.4. *Qualitative observations*

The basic idea of EDI rests on the assumption that all employees have hidden abilities for innovation, and applying an EDI approach implies that all employees are perceived as 'innovation capital' or 'innovation assets' (Kesting and Ulhøi, 2010). As described in Onarheim (2011), EDI can additionally create job satisfaction and be a strong motivational factor for the participants, as they feel involved in important company decisions. Furthermore, EDI can play an essential role in stakeholder management, as decision-makers are involved throughout the innovation process. In the EDI project underlying this study, the initiators, the core team and the participants expressed great satisfaction with the process in terms of project layout, motivation, commitment and output. In the feedback sessions in the EDI project, the most frequent topic was the positive impact the project had on the motivation of those involved (Onarheim 2011).

In the open-ended questions in the online survey used for this study, the employees were given the opportunity to state their opinion of the EDI project and the distributed voting as such. These responses confirmed some of the basic assumptions for the survey, that the ideas were considered good and relevant and presented in an 'equal' way. Furthermore, they confirmed the importance of the EDI project and the survey as motivating, inspiring and exciting for the participants, with several requests for using such broad involvement of employees in other projects.

6. Discussion

This research attempted to provide a first examination of utilising distributed voting schemes as a way to analyse executive selections in the early stages of engineering design. It compared employee votes with the ideas selected by an executive team and investigated the validity of employee voting behaviour and the biases related to such voting. Four hypotheses were tested, and the results indicate that while distributed voting among employees does suffer from potential biases, such as ownership biases and biases towards selecting visually complex products, it is possible to overcome them by building on the principles of WotC. Some consistency in picking the most promising ideas could be found between the employee sample and the executive team, with 3 shared picks in the top 4 and 4 shared picks in the top 6, indicating some measure of validity in the selection method. Still, the executive top 12 picks shared less than 50% with the employee top 12 picks. This shows how some ideas disregarded by the executives were preferred by the employees, and vice versa, even when controlled for the related biases. This indicates that employee ratings can offer executive decision-makers a valuable contribution to a more nuanced picture when deciding what ideas are to be developed further. The results show that it may be worthwhile to further explore whether distributed voting schemes based on WotC could be usable supplements to the standard selection methods of either individual decision-making based on a set of selection criteria or group-based discussion leading to consensus.

The results indicate that in a case such as this, with approximately 100 appropriate ideas to screen and raters who are highly familiar with the problem at hand, reliable measures of idea selection could be expected to be obtained with as less as 30 people making selections, as calculated in the section on inter-rater agreement with a variant of the Spearman Browne Prophecy formula.

The results documented that while it did appear that visual complexity served as a heuristic for determining the idea potential for advancement, it was only the inexperienced employees who seemed to be utilising this heuristic. Thus, a way to overcome the tendency to pick ideas that appear visually complex is to rely on experienced raters more than on inexperienced ones. However, it should be noted that the correlation between the experienced and inexperienced employee groups was quite high, and therefore, although the visual complexity bias did impact on the validity of the results, the impact was somewhat modest in size across the sample. It should be stressed, though, that the present field experiment utilised ideas that were illustrated by a professional designer, thus making the visuals appear somewhat homogeneous from the outset. If a more heterogeneous set of images is to be evaluated in other settings (e.g. if the different respective idea generators themselves have drawn the visuals), then it can be expected that the variation in visual complexity would rise, along with the potential for the effect of the visual complexity bias. As such, if inexperienced raters are utilised in distributed voting schemes, it seems advisable to control for visual complexity of the ideas. While the current explanation for the relation between visual complexity and selection placed employee experience as a mediator, it should be noted that in so far as it is possible to add diversity to the sample, it is possible that increased diversity (and a larger sample) may also counter the bias for visual complexity. This study has identified employee experience as an important mediator by which to screen the sample, but there may be other ways to counter the visual complexity bias. More research is needed on this point.

The second bias was one of ownership, showing that individuals who had proposed or helped further develop an idea had a much higher likelihood of selecting the idea for advancement than other ideas. Although this bias was exceedingly large at the individual level, it had all but disappeared when aggregating across individuals, as suggested by the WotC principle. In this case, it appears that even though the bias to select own ideas was a potent one, it did not matter at all in the overall results of which ideas should be selected. The reason is probably that each evaluator reported having worked on very few ideas (six, on average), and the sample of raters was random and unbiased compared with the sample of individuals who had helped generate the ideas. As such, there was no consistent bias towards ownership of particular ideas in this experiment. It should be noted, however, that if evaluators had worked on a significant proportion of the ideas, and particularly, if multiple raters had worked on the same ideas, then this is likely to provide significant biases. It seems relevant to warn future implementers of employee voting in EDI idea selection to test whether evaluators consistently have a bias towards ownership of particular ideas. In case multiple raters have worked on a large proportion of the ideas, or there is a danger of a skewed or biased sample of raters in terms of idea ownership, it would be advisable to remove ratings of 'own ideas', as the bias proved to be quite potent.

Overall, some validity of the employee-based WotC could be found when comparing with a team of executives. Although the correlation between the two was low (i.e. in the 0.3 range), it was significant. More importantly, among the top picks selected for advancement in the two groups, there was some encouragement in that a sizable number (3 of 4 and 5/7 of 12) of the top picks were shared between the employee crowd and the executive team. Furthermore, the executive team and the experienced employees agreed on picking or not picking ideas in 23 of 26 categories of ideas in the present experiment. It is possible that the differences in picks between the executives and the experienced employees were a result of simply selecting two different but similar ideas in the same class of ideas. Again, this holds promise for both the validity and the reliability of the method. It is to be expected that the executive team utilised a broader range of knowledge areas, such as whether the solution was patentable or if something similar already exists on the market,

which may further explain differences between the top picks in the two employee groups differing in experience level. It is, of course, possible that the executive team to some extent could have made less than optimal choices (e.g. due to social processes such as group think or lack of diversity represented in the group) and thus provided a less than perfect benchmark. The choice of a given benchmark always offers the possibility of the benchmark being slightly off target of course, but the problem is potentially larger in the present context, where we do not have objective *post hoc* information about what ideas could have turned in, had they been given the chance to advance to the last gates, and sent onto the marketplace. In theory, we thus do not know (and cannot know) which of the imperfectly correlated selections (by the executive team and the employee sample selections) constitute the better match to the 'objectively' best possible selection. As such, the choice of benchmarking with an executive team, albeit being supported by the track record of executive team, and the case company, may be subject to criticism.

This research may suffer from limitations. For legal reasons, we had to limit the group sample to employees at the company rather than to a completely random one. The findings may be limited to EDI, as would frequently be the case in non-open innovation. Although efforts were made to maximise the diversity of the employee sample as much as possible, the sampling was limited by organisational affiliation. Further research is needed in order to determine whether organisational affiliation constitutes another bias in the results when compared with random sampling.

Although more research is needed before firm conclusions can be drawn, some recommendations can be extracted from this research. The design of this study does not allow for suggesting WotC as an alternative to other kinds of idea selection since this study does not hold proof of an absolute improvement in idea quality in WotC over other selection methods. In utilising an employee sample in idea selection, it need not, however, be a matter of resolving all problems concerning idea selection, but rather be about supplementing other types of methods. Idea evaluations are crucial in innovation, as argued throughout the paper, and this means not just selecting the right ideas (true positives), but also avoiding selecting the wrong ones (false positives) and avoiding not selecting the right ones (false negatives). Executive teams or individuals operating as gatekeepers are not perfect in their evaluations; they need to make decisions, but they may overlook quality ideas and/or focus on the wrong ones. Distributed voting schemes based on WotC can serve as a supplement to executive idea evaluation, ensuring that unselected executive ideas that ring true to a crowd may be given a second glance before being discarded. Conversely, ideas selected by gatekeepers may also receive a second glance, in so far as the crowd considers them a no-go, in order to establish that organisational resources are not wasted in pursuing these ideas further. This argues for utilising employee voting in addition to other forms of idea screening, in order to ensure that the top picks are indeed the best ones and that executive groups or individuals basing their ratings on inflexible criteria are not inadvertently leaving out good choices (or selecting bad ones) to advance to later stages in product development. In applying distributed voting as a supplement, in so far as variety exists in the visual complexity of ideas, the selections of the raters should be weighed by their experience. Furthermore, if the employees generating the ideas are the same individuals as those rating the ideas, care should be taken to ensure that the selected sample of raters is a random and unbiased one to ensure that ownership biases are avoided. Finally, of course, enough raters should be used to ensure a reliable selection.

Acknowledgements

This study was made possible due to the support and engagement of Coloplast A/S colleagues and company representative Michael Holm Hansen. This work was supported by the Initial Training Network 'Marie Curie Actions', funded by the FP 7 – People Programme with reference PITN-GA-2008-215446 entitled 'DESIRE: Creative Design for Innovation in Science and Technology'.

References

Armstrong, J.S., ed., 2001. *Principles of forecasting*. Dordrecht: Kluwer Academic.

Barczak, G., Griffin, A., and Kahn, K.B., 2009. PERSPECTIVE: trends and drivers of success in NPD practices: results of the 2003 PDMA Best Practices Study. *Journal of Product Innovation Management*, 26 (1), 3–23.

Brabham, D.C., 2008. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14 (1), 75–90. doi:10.1177/1354856507084420.

Cho, H. and Schwarz, N., 2006. If I don't understand it, it must be new: processing fluency and perceived product innovativeness. *Advances in Consumer Research*, 33 (1), 319–320.

Christensen, B.T., Kristensen, T., and Reber, R. (under review). Contributions of perceived creativity and beauty to willingness-to-pay for consumer products.

Cooper, R.G., 2001. *Winning at new products: accelerating the process from idea to launch*. 3rd ed. Reading: Perseus Books.

Cooper, R.G. and de Brentani, U., 1984. Criteria for screening new industrial products. *Industrial Marketing Management*, 13 (3), 149–156.

Cooper, R.G. and Edgett, S.J., 2007. *Generating breakthrough new product ideas: feeding the innovation funnel*. Toronto, Canada: Product Development Institute.

Cooper, R.G., Edgett, S.J., and Kleinschmidt, E.J., 2002. Optimizing the stage–gate process: what best-practice companies do. *Research Technology Management*, 45 (5), 21–27.

Crawford, M. and Di Benedetto, A., 2006. *New products management*. Boston, MA: McGraw Hill.

Faure, C., 2004. Beyond brainstorming: effects of different group procedures on selection of ideas and satisfaction with the process. *Journal of Creative Behavior*, 38 (1), 13–34.

Galton, F., 1907. Vox Populi. *Nature*, 75 (1949), 450–451.

Hogarth, R.M., 1978. A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21 (1), 40–46.

Joshi, A.W. and Sharma, S., 2004. Customer knowledge development: antecedents and impact on new product performance. *Journal of Marketing*, 68 (4), 47–59.

Kahneman, D., Knetsch, J.L., and Thaler, R.H., 1991. Anomalies: the endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives*, 5 (1), 193–206.

Kamp, G. and Koen, P.A., 2009. Improving the idea screening process within organizations using prediction markets: a theoretical perspective. *The Journal of Prediction Markets*, 3 (2), 39–64.

Kesting, P. and Ulhøi, J.P., 2010. Employee-driven innovation: extending the license to foster innovation. *Management Decision*, 48 (1), 65–84.

LaComb, C.A, Barnett, J.A., and Pan, Q., 2007. The imagination market. *Information Systems Frontiers*, 9 (2/3), 245–256.

Moreau, C.P., Lehmann, D.R., and Markman, A.B., 2001. Entrenched knowledge structures and consumer response to new products. *Journal of Marketing Research*, 38 (2), 14–19.

Onarheim, B., 2011. Using a company Brainstorm for employee-driven innovation: a case study. *Design Principles and Practices: An International Journal*, 4 (6), 347–354.

O'Quin, K. and Besemer, S.P., 1989. The development, reliability, and validity of the revised Creative Product Semantic Scale. *Creativity Research Journal*, 2 (4), 267–278.

Prahalad, C.K. and Ramaswamy, V. 2004. Co-creation experiences: the next practice in value creation. *Journal of Interactive Marketing*, 18 (3), 5–14.

Reid, S.E. and de Brentani, U., 2004. The fuzzy front end of new product development for discontinuous innovations: a theoretical model. *Journal of Product Innovation Management*, 21 (3), 170–184.

Rietzschel, E.F., Nijstad, B.A., and Stroebe, W., 2006. Productivity is not enough: a comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental Social Psychology*, 42 (2), 244–251.

Rietzschel, E.F., Nijstad, B.A., and Stroebe, W., 2010. The selection of creative ideas after individual idea generation: choosing between creativity and impact. *British Journal of Psychology*, 101 (1), 47–68.

Simmons, J.P., *et al.*, 2011. Intuitive biases in choice versus estimation: implications for the wisdom of crowds. *Journal of Consumer Research*, 38 (1), 1–15.

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

Surowiecki, J., 2004. *The wisdom of crowds*. New York: Doubleday.

von Hippel, E., 1986. Lead users: a source of novel product concepts. *Management Science*, 32 (7), 791–805.

Young, M.E. and Racey, D., 2009. Judgments of creativity as a function of visual stimulus variability. *Empirical Studies of the Arts*, 27 (1), 89–107.