IDEA SCREENING IN ENGINEERING DESIGN USING EMPLOYEE-DRIVEN WISDOM OF THE CROWDS

Balder Onarheim and Bo T. Christensen
Copenhagen Business School, Denmark

ABSTRACT
The paper investigates the question of screening ideas in the 'fuzzy front end' of engineering design, examining the validity of employee voting schemes and related biases. After an employee-driven innovation project at a major producer of disposable medical equipment, 99 ideas were to be screened for further development. Based on the concept of 'wisdom of the crowds', all ideas were individually rated by a broad selection of employees, and their choices of ideas and idea categories compared to those of a small team of senior marketers. The study also tested for two biases: visual complexity and endowment effect/ownership of ideas. The study shows that the crowd wisdom of employees significantly correlates with the preferences of the marketing team: overall, in top 12 selected ideas and in choice of idea categories. This match increases when including only the ratings of the most experienced employees. The experienced employees also proved to be less affected by visual complexity in the ideas presented. The endowment effect was potent in that every employee proved to be more likely to select their own ideas over others, but this effect disappeared when aggregating across the crowd of employees.

Keywords: Evaluation of creative ideas, creativity, idea evaluation, idea screening, engineering design, fuzzy front end, innovation, evaluator experience, new product development, wisdom of the crowds, employee-driven innovation, idea filtering.

1. INTRODUCTION
Idea evaluation and -selection, particularly during the early stages of engineering design, is notoriously fraught with uncertainty. While the idea generation process in innovation has been examined in quite some research, the early evaluation process (how are ideas to be evaluated, who is evaluating and by which criteria) has not been the subject of much research, even though it represents an important element in the 'fuzzy front end' of new product development [1]. A pool of new ideas is alone an insufficient condition for innovation, as the importance lies equally in the recognition and selection of the best ideas [2]. Having a number of good ideas generated does not matter if they are not picked out for progression to later product development stages. Such selection is therefore a major challenge in the fuzzy front end of new product development [3]. This early selection process is a critical, difficult and complex task [4]. Particularly when, as is frequently the case, multiple ideas have been generated and company resources only permit very few of these to be turned into actual company projects. How are we to select the right ideas for progression? Typical solutions involve a selected few executives making the decision, or a small panel basing their evaluation on inflexible criteria, such as what Cooper describes as 'must have' and 'should have' [5, 6]. 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. 5]. Such stage-gate model theories tend to focus on the criteria to apply 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, portfolio management of the range of ideas that should enter into R&D projects has been examined [7]. An overview of previous research investigating methods for filtering and evaluation of new product ideas can be found in [8].
As an alternative approach to the use of expert teams to select the best ideas, this study is investigating the use of the wisdom of the crowds (WotC) in the fuzzy front end, by asking multiple employees to vote for the best ideas after a company wide brainstorm. As pointed out by [9]: *"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"*. Such distribution of decisions is in line with

concepts like employee-driven innovation (EDI) [10, 11], idea sourcing [12] and idea markets [3], building on the idea of WotC [9]. To date, not much research has actually examined the validity of such distributed voting schemes for selecting ideas. Furthermore, it seems relevant to suggest that voting by 'lay' people, or a broad selection of employees, might be fraught with potential biases.

1.1. Selecting the right idea

In addition to the above-mentioned challenges with selecting ideas in the fuzzy front end, idea selection in general is considered a notoriously difficult process. Not only do many companies lack a coherent or formal process for selecting ideas [13], studies show that people perform very poorly at selecting their own most creative ideas as well [14, 15]. Reitzschel et al. even found [15] that the ideas selected for their creativity in some cases was no better than randomly sampling the pool of generated ideas! Clear criteria for selection has been pointed out by some scholars as an important step towards improvement of the quality of selected ideas [2, 5], 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 [4] makes it challenging to know what criteria to focus 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 the issue of establishing clear criteria for product development processes [6].

An alternative to establishing distinct and clear criteria in creative judgment is to rely on holistic judgments of products without explicating the dimensions to be rated. The Consensual Assessment Technique is one such approach [16]. In this approach, a number of independent ideas/products are evaluated for their level of creativity by independent and appropriate judges. In the consensual assessment technique, reliability is assumed to basically correspond to construct validity [16]. Research has repeatedly shown that it is possible to reliably estimate the level of creativity in products in such an experimental framework [17]. The consensual assessment technique has later been extended to also being able to handle ideas generated in nonparallel (i.e., non experimental) settings [18]. While 'appropriate' judges originally entailed 'experts', it has been shown that in some cases less experienced raters are also able to provide reliable estimates of creativity [17], although see [19] for results indicating low reliability in novice judges. We wanted to extend these findings to real-world decisions in engineering design concerning the picking of promising ideas to be made into projects (rather than the more restricted question of evaluating the level of creativity in the product). Is it possible to utilize a crowd of more or less randomly picked employees as a valid source of ideas selection? Furthermore, we wanted to check whether employees who participated in EDI workshops (i.e., who had gained some knowledge of the innovation challenge through their own solution attempts during the workshops) could serve as 'appropriate' judges, despite their varying degrees of background experience. Finally, we wanted to add ecological validity to the research design, by utilizing real-world engineering design problems and ideas from a large international company working in medical plastics.

1.2. Wisdom of the crowds

The WotC hypothesis predict that the independent judgment 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 [9, 20]. For example, Francis Galton [21] 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, by experts or in group decisions. The hypothesis is derived from mathematical principles, in that a crowd's judgment comprises signal-plus-noise, and averaging across judgments will then cancel out the noise while extracting the signal [9, 22]. The conditions for occurrence of the WotC are that 1) the crowd is knowledgeable, and 2) individual errors in judgment must not be systematic at the sample level. Systematic errors in judgment can for example occur with restricted diversity on the judging sample or lack of independence amongst the judges.

In an organizational context, it is important to not confuse this use of WotC with a compact internal "market study". An employee-driven WotC focus on which ideas the employees believe are important for the company to continue to develop on, and thus draws on both internal organizational knowledge, external market knowledge, and product focused technical knowledge. The employees used for rating

is not necessary customers, and they are not asked to estimate the market potential for each idea, but rather to focus on their general understanding of which ideas are worthy of further development in the present context. It's important to try to determine whether some form of bias may be leading the crowd to make erroneous or poor decision. In the present paper we examine the potential impact on the WotC by two sources of bias, as well as two ways to overcome them: visual complexity and endowment effects.

1.3. 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 [23, 24], and recent research has shown how increasing complexity or lowering visual fluency lead to higher ratings of creativity [25] or product innovativeness [26]. 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 – criteria when estimating which ideas should be allowed to progress though gates in a product development process. This led to the first hypothesis:

H_1: High visual complexity in the presentation of the individual idea leads to more selections of that idea for further development.

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 like the link between visual complexity and creativity. Experts and novices often disagree systematically in their selections of product ideas [27], and results from forecasting studies stress that using several experts instead of one leads to better results [28]. In addition, Cooper [5] has argued for the need for experienced judges in product evaluation in product development gates. Therefore…

H_2: Experienced raters should not rely on visual complexity heuristics to the same degree as inexperienced raters.

1.4. Endowment / ownership effects

When ideas have been generated, it has been shown that the creators or contributors to the idea generation lead them to hold their own ideas in higher esteem compared to other ideas. In behavioral economics, this has been labeled the endowment effect, whereupon it has been shown that investing time and energy in developing solutions leads you to appreciate that solution more, and owning an object/solution leads to increased feelings of loss when having to let it go [e.g. 29]. Cooper [5] describes 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 [6]. 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.

H_3: Ownership of an idea leads to a higher rate of selection of that idea by individual raters.

There are different ways of trying to counter this well-known bias. Cooper et al. [6] argued that it could be countered through the setup of clear selection criteria ('must have', 'should have') to be implemented rigorously at the gates. However, such clear and rigorous criteria are both hard to formulate unambiguously (which is why so many companies have a problem with implementing them), and further it is extremely difficult to find objective ways to weight these criteria against each other in the selection process [3]. 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 a WotC 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). Thus…

H_4: The ownership bias should disappear in utilizing wisdom of the crowd, 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, we compared the employees ratings 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 [30]. Previous research estimating the validity of idea selection have utilized the same type of 'expert team' measure against which the wisdom of the crowds could be measured, and have generally found somewhat low levels of validity ranging in the .10 to .47 [3, 31].

2. METHODS

The present research attempted to provide a first examination of biases and validity of employee voting behavior in company idea selection.

The research design was a real-world field study conducted in a major international company dealing in disposable medical equipment. The base for the 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 workshops, generating a pool of 99 distinct ideas described in writing and drawings and sorted in 26 different categories. As pointed out by [12], such a large number of contributors with diverse skills can enhance the chances of finding a truly innovative idea. It is important to notice that all ideas were subject to an ongoing screening in the workshops, a process that should ensure that all the 99 ideas can be considered of a satisfactory quality. Furthermore, the grouping of the ideas into categories was an emphasized part of the workshops, ensuring relevant and homogeneous categories. We therefore use categories as a measure in the study, as we assume that ideas within one category are not considered fundamentally different. For a comprehensive description of the generation of the process, the expert selection and the rationale behind the lay out, see [11].

2.1. 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 bullet point fashion. Furthermore, the drawings were redrawn by a professional designer, resulting in a catalogue presenting all 99 ideas in a standardized manner. Each idea was, as illustrated in figure 1, presented on a horizontally oriented A4 page, one half of the page with a short description of the idea in text, including a list of "Customer benefits" and "Coloplast 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 we are therefore unfortunately restricted from sharing the actual ideas.

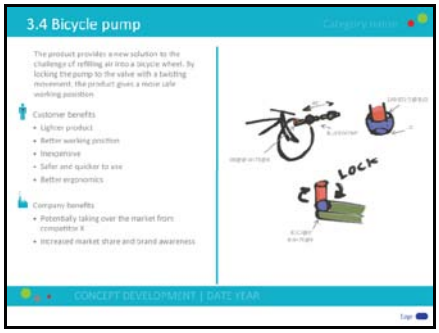

Figure 1

2.2. Measure of complexity

Each idea was rated by an independent researcher unaware of the hypotheses of this article for visual, textual and benefit complexity. Visual complexity was counted as the number of separate drawings made to visualize 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, i.e. with more than 6 letters). Benefit complexity was calculated simply as the number of bullet points describing the benefits of the idea for the company and the users.

2.3. Product evaluation

Two groups of company employees independently evaluated each idea. As part of the EDI process, company executives selected an expert team consisting of 7 handpicked senior marketers representing 4 national subsidiaries and the main office. Using such expert teams to select ideas for advancement is a usual way to filter ideas for new product development at the company. This expert team was gathered for a full day workshop where all the ideas were assessed, with the criterion of finding the best ideas suited for further development. The group discussed what they considered important 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 expert team's evaluation served as the standard against which the wisdom of the employee crowd was compared. This was not done in order to claim that the expert 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 experts are the ideas that actually will be realized. When ideas in product development are screened and some discarded, there exists no objective knowledge about what would have been the best ideas [30], thus we selected the expert team as the best measure in order to estimate the validity and accuracy of the wisdom of the crowds.

In addition to the expert team, the employees contributing with 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 deployed at the company. The employee crowd evaluating the products was a group of 35 employees (16 female, 19 male, mean age 42) from a variety of job functions and company departments, who had taken part in the workshop. They represented involvement from 11 departments, had a mean of 8 years company experience (range 0 to 24) and 4 years experience working in the product domain in question (range 0 to 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 catalogue the participants were first 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 99 ideas one by one, randomized for ordering across participants, and answered the two questions. Furthermore, information about level of expertise of the individual employee was obtained.

3. RESULTS

The mean number of times an idea was selected of the 35 raters was 14.6 (STD 7.1, ranging from 1 to 32). 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 (STD 12.5, ranging from 15 to 67 ideas).

The agreement among judges was satisfactory. ICC (two-way for consistency) among the 35 raters for their selections was .87. It is possible to calculate how many evaluators would have been needed in

order to reach a satisfactorily high level of agreement (ICC >.8) using a variant of the Spearman Browne Prophecy formula

$$m = \frac{\rho^*(1-\rho_I)}{\rho_I(1-\rho^*)}$$

where m is result to be rounded to the next highest integer, ρ^* is an aspiration level, and ρ_L is a reliability estimate, typically either ICC(2,1) or ICC(3,1). For an experimental setup like this (with 99 individual ideas to be rated, and random judges), we should expect that to replicate the high level of agreement, 30 individual raters should have sufficed.

To investigate H_1, whether idea complexity biased subjects towards selection, we standardized the three kinds of complexity (textual, visual and benefit) and averaged across them, to generate a total complexity measure. A linear regression of whether the total complexity measure predicted selection of the individual ideas producing an adjusted R^2 of .048 ($F(1, 98)=5.98$, $p<.02$) with total complexity being a significant predictor ($\beta=.24$, $t(98)=2.45$, $p<.02$), showing that idea complexity did predict selection.

To further examine H_2, whether employee expertise moderated the idea complexity bias, we divided the employees 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 experience in the product domain. The experienced group (N=15) had a mean of 14 years company experience and 8 years domain experience, and the inexperienced group (N=20) had a mean of 3 years company experience and 2 years 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 .058 ($F(3, 89)=2.89$, $p<.04$). Visual complexity ($\beta=.27$, $t(98)=2.69$, $p<.01$) was significant, while textual ($\beta=.12$, $t(98)=1.20$) and benefit complexity ($\beta=.07$, $t(98)=.72$) were nonsignificant 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 .076 ($F(3, 89)=3.52$, $p<.02$). Benefit complexity ($\beta=.23$, $t(98)=2.33$, $p<.03$) was significant, while visual ($\beta=.19$, $t(98)=1.88$) and textual complexity ($\beta=.16$, $t(98)=1.54$) were nonsignificant predictors. In comparison, the same regression was run for the selection of the marketing expert team, yielding an adjusted R^2 of .051 ($F(3, 89)=2.64$, $p=.054$). Benefit complexity ($\beta=.23$, $t(98)=2.24$, $p<.03$) was significant, while visual ($\beta=-.05$, $t(98)=-.46$) and textual complexity ($\beta=.18$, $t(98)=1.75$) were nonsignificant predictors. The results indicate that while both the marketing expert 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 (number of visual images shown).

To examine H_3, whether having worked on an idea biased evaluators towards selecting that idea, we calculated the proportion of ideas selected for the ideas the evaluator had or had not worked on respectively. Thirteen evaluators did not report having worked on any of the ideas, even though they had been present in at least one idea generating workshop. A paired t-test showed a significant difference (paired-$t(21)=10.34$, $p<.001$), with a mean probability of picking an idea the evaluator had worked on of .81, with .40 for ideas not reported to have worked on. The effect was so potent that for every evaluator there was a higher average proportion of picks for ideas that had been worked on compared to ideas not worked on.

To estimate (H_4) whether expertise mediated the ownership effect identified in H_3 when utilizing the wisdom of the crowds, we used the expert marketing team as a benchmark. In overall the mean of the employee crowd selections correlated ($r=0.32$, n=99, $p=.001$) with the selection of the marketing 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 expert team, and the crowd. To estimate this, the 12 ideas (paralleling the 12 picks of the marketing team) with the most votes were considered 'picks of the crowd'. Furthermore, given that the ideas were categorized in 26 categories by overall topic, it was possible to also estimate how many of the general categories that the expert marketeers and the crowd had agreed on selecting/not selecting ideas from. Among the top 12

marketing picks, 5 of the ideas were also ranked in the top12 by the employees (Cohen's κ =.34), and of the 26 categories of ideas in the pool, the two groups agreed in their picking/not picking an idea from a category in 21 of the categories (Cohen's κ =.56).
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 top12 actually became a top16 (experienced employees) and top14 (inexperienced employees) in order to accommodate ideas with equal scores.
In comparison, the experienced group of employees correlated (r=0.33, n=99, p=.001) with the marketing team, shared 7 of the top12 picks (Cohen's κ =.42) and agreed on picking/not picking 23 of the categories (Cohen's κ =.75); while the inexperienced group correlated (r=.29, n=99, p=.004), shared 5 of picks in the top12 (Cohen's κ =.29) and agreed on picking/not picking 21 of the categories (Cohen's κ =.59). The experienced and inexperienced employee groups selections correlated (r=.78, n=99, p=.001), shared 11 of picks in their top14/16 (Cohen's κ =.69) and agreed on picking/not picking 20 of the categories (Cohen's κ =.51). As such, some significant gains appear to result from choosing experienced employees as evaluators – particularly in the shared picks in the top12, and in choosing the general categories from which ideas would be selected in this case. To further illustrate the ratings of the experienced/inexperienced raters, we calculated the proportion of the two groups of employees (high/low experience) that had selected the 12 ideas picked by the marketing expert team, as shown in figure 2.

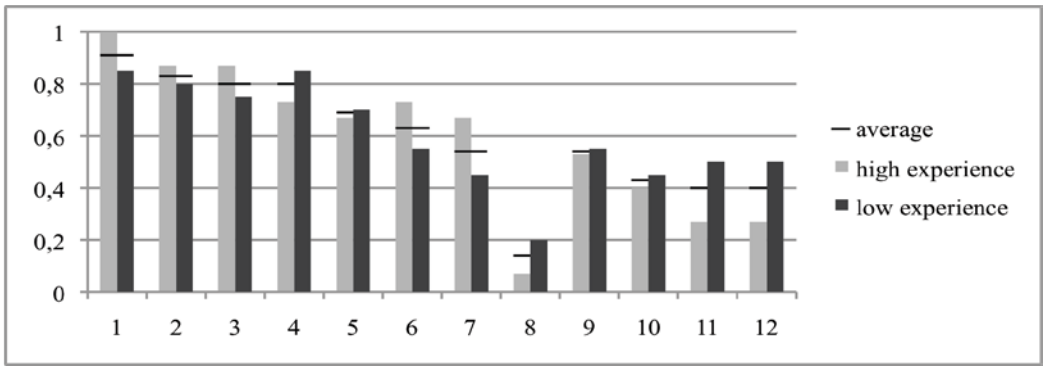

Figure 2

The figure shows how half of the ideas picked by the marketing team were selected by more than 60 % of employees (1-6). Five of the 12 ideas were preferred to a larger extend by the high experience group (1, 2, 3, 6, 7), while seven of the (on average) less popular ideas were preferred to a larger extend by less experienced raters (4-5, 8-12). Especially idea 8, 11 and 12 stand out as having a lower rating by the high experience group, and these are also ideas with the lowest overall proportion picks. Idea 8 is the outlier, with only 14 % picks in the average employee ratings, and 7% picks amongst high experienced raters.

To estimate whether the ownership bias could be countered we excluded all data on ideas the individual evaluator had worked with. The correlation of this new mean measure to that of the marketing team did not show any differences: the correlation was still (r=.32, n=99, p=.001) and there were still 5 shared picks in the top 12 (Cohen's κ =.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.

4. SUMMARY AND DISCUSSION

The present research attempted to provide a first examination of the validity of employee voting behavior for idea selection in engineering design, and the biases related to such voting. Four hypotheses were tested, and the results indicate that while employee-driven WotC does suffer from potential biases, such as ownership biases and biases towards selecting visually complex products, it is possible to overcome them using WotC. Furthermore, some consistency in picking the most promising ideas could be found between employee crowds and an expert marketing team, indicating some measure of validity in the selection method.

The results indicate that in a case like this, with approximately 100 ideas to screen and raters who had worked on the problem to be solved in lengthy workshop sessions, reliable measures of idea selection could be expected to be obtained with as little as 30 people making selections. This informs the literature on the Consensual Assessment Technique in generalizing the technique to not only covering ratings on the level of creativity in the products or ideas themselves, but also their potential for advancement in product development stages. As such, the results show that it may be worthwhile to further explore whether WotC could be a usable alternative or supplement 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 documented that while it did appear that visual complexity served as a heuristic for determining idea potential for advancement, it was only the inexperienced employees who seemed to be utilizing this heuristic. Thus a way to overcome the tendency to pick ideas that appear visually complex is to rely on experienced raters more than inexperienced ones. However, it should be noted that the correlation between the experienced and inexperienced employee crowds 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 crowd. It should be stressed, though, that the present field experiment utilized ideas that were illustrated by a professional designer, thus making the visuals appear somewhat homogenous 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 utilized in WotC, it seems advisable to control for visual complexity of the ideas.

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, compared to other ideas. Although this bias was exceedingly large at the individual level, it had all but disappeared when aggregating across individuals in the WotC technique. In this case, it appears that even though the bias to select own ideas was a potent one, it did not matter at all to the overall results of which ideas should be selected. The reason is probably that each evaluator reported having worked on very few ideas (6 ideas on average), and the sample of raters was random and unbiased compared to 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 have worked on a significant proportion of the ideas, and particularly if multiple raters have worked on the same ideas, then this is likely to provide significant biases. It seems relevant to warn future implementers of wisdom of the crowds 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 and 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 is quite potent.

Overall, some validity of the employee-driven WotC technique could be found when comparing to an expert team of senior marketers. 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 (5-7 of 12) of the top picks were shared between the employee crowd and the expert marketing group. Furthermore, the expert marketing group and the experienced employees agreed on picking or not-picking ideas in 23 out of 26 categories of ideas in the present experiment. It is possible that the differences in picks between the expert marketing team and the experienced employees was 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 reliability of the method. Further research is needed in order to qualitatively analyze each case where the expert team of marketers and the experienced employees differed. It is to be expected that the expert team utilized 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 groups. It is of course also possible that the expert marketing team to some extent made less than optimal choices, and thus provided a less than perfect benchmark (e.g., due to social processes such as group think or lack of diversity represented in the group).

Although more research is needed before firm conclusions can be drawn, some recommendations can be extracted from the present research: Our study design does not allow us to suggest WotC as an alternative to other kinds of idea selection since we do not have proof of an absolute improvement in

idea quality or creativity in WotC over other selection methods. However, it is notable that the WotC technique was able to provide reliable picks from only 30 respondents. At least it is possible to utilize WotC as a supplement to other forms of idea screening, in order to ensure that the top picks are indeed the best ones and that expert groups or individuals basing their ratings on inflexible criteria are not inadvertently leaving out good choices to advance to later stages in product development. In applying WotC as a supplement, in-so-far as variety exists in the visual complexity of ideas, raters selections should be weighed by rater experience. Further, if the employees rating the ideas are the same individuals as the employees selecting 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: The writing of this paper was partly 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

[1] Reid S.E. and de Brentani U. The Fuzzy Front End of New Product Development for Discontinuous Innovations: A Theoretical Model. *Journal of Product Innovation Management*, 2004, 21(3), 170-184.

[2] Rietzschel E.F., Nijstad B.A. and Stroebe W. The selection of creative ideas after individual idea generation: choosing between creativity and impact. *British journal of psychology*, 2010, 101, 47-68.

[3] Soukhoroukova A., Spann M. and Skiera B. Sourcing, Filtering, and Evaluating New Product Ideas: An Empirical Exploration of the Performance of Idea Markets. *Journal of Product Innovation Management*, 2010, 1, 1-33.

[4] Cooper R.G. and de Brentani U. Criteria for Screening New Industrial Products. *Industrial Marketing Management*, 1984, 13(3), 149-156.

[5] Cooper R.G. *Winning at new products: Accelerating the process from idea to launch, 3rd edition*, 2001 (Perseus Books, Reading).

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

[7] Cooper R.G. and Edgett S.J. *Generating breakthrough New Product Ideas: Feeding the Innovation funnel*, 2007 (Product Development Institute, Toronto).

[8] Crawford M. and Di Benedetto A. New Products Management, 2006 (McGraw Hill, Boston).

[9] Simmons J.P., Nelson L.D., Galak J. and Frederick S. Intuitive Biases in Choice versus Estimation: Implications for the Wisdom of Crowds. *Journal of Consumer Research*, 2010, 38, 000.

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

[11] Onarheim B. Using a Company Brainstorm for Employee-Driven Innovation: A case study. *Design Principles and Practices; An International Journal*, 2011, 4(6), 347-354.

[12] Joshi, A.W. and Sharma S. Customer Knowledge Development: Antecedents and Impact on New Product Performance. *Journal of Marketing*, 2004, 68(4), 47-59.

[13] Barczak G., Griffin A. and Kahn K.B. PERSPECTIVE: Trends and Drivers of Success in NPD Practices: Results of the 2003 PDMA Best Practices Study. *Journal of Product Innovation Management*, 2009, 26(1), 3-23.

[14] Faure C. Beyond brainstorming: Effects of different group procedures on selection of ideas and satisfaction with the process. *Journal of Creative Behavior*, 2004, 38, 13–34.

[15] Reitzschel M.A., Nijstad B.A. and Stroebe W. Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental Social Psy*chology, 2006, 42, 244-251.

[16] Amabile T. Social Psychology of Creativity: A consensual Assessment technique. *Journal of Personality and Social Psychology*, 1982, 43(5), 997-1013.

[17] Hennessey B.A. and Amabile T. *Consensual Assessment*. In Runco M.A. and Pritzker S.R. (eds.) *Encyclopedia of Creativity*, 1999 (Academic Press, San Diego).

[18] Baer J., Kaufman J.C. and Gentile C.A. Extension of the consensual assessment technique to non parallel creative products. *Creativity Research Journal*, 2004, 16(81), 113-117.
[19] Kaufman J.C., Baer J., Cole J.C. and Sexton J.D. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 2008, 20(2), 171-178.
[20] Surowiecki J. *The Wisdom of Crowds*, 2004 (Doubleday, New York).
[21] Galton F. Vox Populi. *Nature*, 1907, 75, 450-51. Quoted in Surowiecki J. *The Wisdom of Crowds*, 2004 (Doubleday, New York).
[22] Hogarth R.M. A Note on Aggregating Opinions. *Organizational Behavior and Human Performance*, 1978, 21(1), 40–46.
[23] O'Quin K. and Besemer S.P. The development, reliability, and validity of the revised Creative Product Semantic Scale. *Creativity Research Journal*, 1989, 2(4), 267–278.
[24] Young M.E. and Racey D. Judgments of creativity as a function of visual stimulus variability. *Empirical studies of the arts*, 2009, 27, 89-107.
[25] Christensen, Ball & Reber (under preparation). Fluency effects of judgments of creativity and beauty.
[26] Cho H. and Schwarz N. If I don't understand it, it must be new: Processing fluency and perceived product innovativeness. *Advances in Consumer Research*, 2006, 33, 319-320.
[27] Moreau C.P., Lehmann D.R and Markman A.B. Entrenched knowledge structures and consumer response to new products. *Journal of Marketing Research*, 2001, 38(2), 14-19.
[28] Armstrong J.S. (ed.) *Principles of forecasting*, 2001 (Kluwer Academic Publishers, Dordrecht).
[29] Kahneman D., Knetsch J.L. and Thaler R.H. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *The Journal of Economic Perspectives*, 1991, 5(1), 193-206.
[30] Kamp G. and Koen P.A. Improving the idea screening process within organizations using prediction markets: A theoretical perspective. *The Journal of Prediction Markets*, 2009, 3(2), 39-64.
[31] LaComb C.A, Barnett J.A. and Pan Q. The Imagination Market. *Information Systems Frontiers*, 2007, 9(2/3), 245-256.

Contact: Balder Onarheim
Copenhagen Business School
Department of Marketing
Solbjerg Plads 3D, 3.38; 2000 Copenhagen
Denmark
Tel: +45 50 373 555
Email: balder@onarheim.com
URL: http://onarheim.com

Balder Onarheim is a PhD Fellow at Copenhagen Business School (www.cbs.dk), and holds a master in Industrial design from Oslo School of Architecture and design (www.aho.no). His main interest is creative processes in complex design, and the role of constraints in shaping creative environments. Bo T. Christensen is an Associate Professor at Copenhagen Business School, and is a cognitive psychologist by training. His theoretical focus is on creative cognitive processes such as analogy, simulation and incubation.