Conceptualizing the Idea Selection Problem: Building on Insights from a large-scale Innovation Contest

Martin Stoetzel

University of Erlangen-Nuremberg, Chair of Information Systems III, 90403 Nuremberg,
E-Mail: martin.stoetzel@fau.de

Martin Wiener

University of Erlangen-Nuremberg, Chair of Information Systems III, 90403 Nuremberg,
E-Mail: martin.wiener@fau.de

Abstract

Web-based innovation contests as a means to engage a community of innovators have become a popular instrument in the past years. While several studies reported on successful idea selection in their cases, we actually were confronted with unexpected difficulties. In order to gain insights into the idea selection problem in large-scale innovation contests with 100 or more ideas, we first develop a conceptual idea selection model and then run Monte-Carlo simulations for 810 idea selection scenarios. Our results not only confirm "common sense" understanding (e.g. more raters better than few, consistent rating better than inconsistent): By using different value distributions we are also able to quantify the effects of certain design parameters in terms of rating performance, which is a new approach compared to previous studies on idea rating and selection. Our findings could thereby help scholars and practitioners to optimize the idea selection process in the future.

1 Introduction

This study originates from our involvement in organizing a large-scale web-based innovation contest with students. Since 2010 this innovation contest is part of an undergraduate course at a major German university, with more than 1,000 students participating in this contest each winter term. Contest participation makes 25% of the final grade for the course, thereby encouraging all students to contribute novel and relevant ideas for a defined challenge. In the winter term 2012/13, the contest was organized in two phases: A concept development phase and a subsequent evaluation phase. The evaluation was done by two distinct groups: The students themselves evaluated concepts from other students, and in parallel a jury consisting of 27 research assistants from the organization committee evaluated the ideas. Table 1 shows some key figures from the contest run in 2012/13:

Participants	Ideas	Participant ratings	Jury ratings	Ratings per idea
1,445 in 310 teams	310	14,438	878	~49

Table 1. Students innovation contest 2012/13

Other than in previous years, the contest in 2012/13 was for the first time run in collaboration with an external partner, the City Council. Together with representatives from the City Council we defined the contest challenge as "developing ideas and concepts for innovative digital services to be used by citizens as well as tourists and commuters". The students were randomly grouped into teams of 5 and they developed in total 310 ideas between Oct-31-2012 and Nov-28-2012. For the evaluation phase, we applied a carefully designed multi-criteria evaluation approach for obtaining the idea ratings. Very similar to [4] and [17], the following eight criteria had been defined and agreed with the city council: "Problem orientation", "elaboration", "novelty", "user value", "user acceptance", "marketing potential", "technical feasibility" and "economic feasibility". Each idea was rated by ~46 students and also by 3 members from the jury. The evaluation was done for all criteria on a 7-point Likert scale.

The overall rating was then calculated as the sum of the individual scores and divided by the number of raters per idea. Building upon earlier studies, we were expecting that the criteria and the evaluation approach would be adequate to identify the best ideas. However, after comparing the jury ranking with the students ranking, we found that their evaluation was quite diverse. Figure 1 shows the jury ranking compared to the students ranking – the ideas on the x-axis are sorted by the jury ranking from best (1) to worst (310). A Spearman rank correlation test produced a correlation coefficient $\rho = 0.27$ which is significant at 0.01 level; but if we for instance compare agreement on the best 25 ideas, we only find 6 ideas which are both in the top-25 from the jury and in the top-25 from the students.

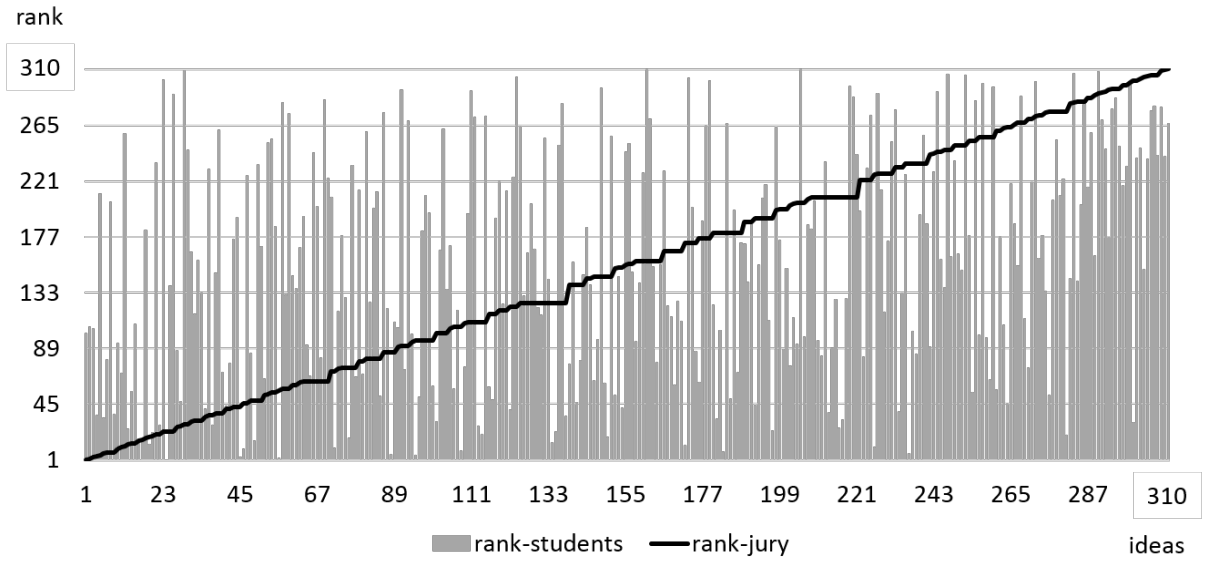

Figure 1. Idea ranking from students and jury compared (sorted by jury ratings)

Although past research on innovation contests has presented a number of insights with regards to idea selection, we found two limitations which we want to address in this study: (1) Very little is known about the idea selection problem in large-scale innovation contests; studies which discuss idea selection in detail have predominantly built on contests with a smaller set of ideas; and (2) those studies which discuss idea selection uniformly report that the process and the result have been satisfactory, i.e. we did not find any study which reported about unsatisfactory evaluation results and measures for avoiding or mitigating such situations. Therefore, our objective is to conceptualize the idea selection problem for large-scale innovation contests and to derive insights into relevant design parameters. Our findings are also meant to help the obviously growing number of practitioners who engage the "innovation community" and who are confronted with the task of selecting the best ideas.

2 Theoretical Background

2.1 Idea Selection in Innovation Contests

In the past years, web-based innovation contests have been conceptualized from various perspectives, e.g. the design elements and parameters of innovation contests, motivation of participants, award design, and characteristics of participants (cf. [25] p.10). In the following, we will concentrate on studies which report on *idea evaluation* and *idea selection* in innovation contests or similar settings.

We performed a systematic literature review via the EBSCOhost Business Source Complete database using keywords as "innovation contest", "ideas competition", "innovation challenge" and some more variations. The aim was to find studies which report on innovation contests and the principles for selecting the best ideas. We also applied a backward search (relevant references cited in the found studies) and complemented our search by a forward search (later studies citing our results) via Google Scholar. Interestingly, although web-based innovation contests have become increasingly popular (see the Innovation Contest Inventory database [12] for many examples), we could not find many studies which present detailed insights about idea evaluation and idea selection in innovation contests:

Studies	Contest	Year	Participants	Ideas
[2], [7], [18], [24]	SAPiens	2008	39	57
[19], [24]	mi-adidas & ich	2004	57	82
[16]	Telia Mobile	n.a.	47	251
[8]	WIN Contest	2010	1,198	234
[9]	CEC Shoe Design	2007	~ 400	66
[10]	Ideenschmiede	2011/12	194	56

Table 2. Studies discussing web-based innovation contests (including idea evaluation)

What all these studies have in common is that the selection of the best ideas was ultimately done by a jury of so-called "experts" which are employees of the host organization, in some cases complemented by qualified externals. In addition, a couple of studies report that the idea selection has also been influenced by a "community" evaluation, which often preceded the jury evaluation phase. The community evaluation is done either via binary votes (e.g. idea is seen as good or bad, as in [8]) or by using multi-criteria evaluation approaches (e.g., [18], [24]). Most innovation contests applied multi-criteria ratings which is in line with [22] who in their analysis found that simple rating mechanisms were less effective.

With regards to the evaluation criteria, although different sets of criteria were used in the different cases, the rating criteria were neither contradictory nor completely diverse: "Novelty", "relevance", and "elaboration" commonly form the basis for a more or less elaborate set of criteria in each of the cases. This is actually one of the requirements defined by [1] who introduced the *consensual assessment technique* (CAT) for the evaluation of creative work. All listed studies report to having applied the CAT, at least for the expert evaluation of their ideas.

2.2 Rating Reliability

One key requirement of the CAT is a high degree of inter-rater reliability. [1] suggested that sufficient reliability would be achieved with intra-class correlation coefficients (ICC) greater than 0.7. At this point we shall briefly rethink the rationale for using ICC: Intra-class correlation is especially useful for comparing ratings from more than two raters, and when raters are exchangeable, i.e. they are part of a potentially larger population of raters. In the studies listed in section 2.1, ICC was calculated as measurement for the reliability of the expert ratings regarding single rating criteria and/or overall rating scores. The authors of the studies were consistently satisfied with their ratings: [2] report that ICC coefficients were > 0.7 or slightly below for all criteria which were not excluded after factor analysis. Also [19] found high consensus for each of their criteria with ICC coefficients above 0.7. Consensus among judges was also approved in [9] and [10].

In contrast to these studies, the jury ratings from our contest seem to be significantly worse. Table 3 shows that the jury ICC coefficients for all 8 criteria are significantly below the 0.7 threshold. The students evaluation ICC coefficients on the other hand are much higher, most of them greater than 0.7.

Rating criteria	ICC (1,k) Jury	ICC (1,k) Students
Problem orientation	-0.110	0.574
Elaboration	0.108	0.788
Novelty	0.272	0.900
User value	0.068	0.749
User acceptance	0.220	0.769
Marketing potential	0.178	0.825
Technical feasibility	0.357	0.858
Economic feasibility	0.066	0.691

Table 3. Intra-class correlation coefficients (one-way random) for jury and students

In order to correctly interpret these results, we need to take a deeper look at the ICC calculations. First of all, there are different ICC models to be used: If not all raters evaluate each idea, the ICC (1, r) must be chosen which does not take into account effects of individual raters [23]. If the same raters evaluate all ideas, the commonly used model is ICC (2, r), which treats the raters as sample part of a potentially larger population of raters. Finally if raters consistently rate all ideas and they are not interchangeable, the ICC (3, r) model should be preferred. In our case we used the ICC (1, r) model because the evaluation task was shared among a larger set of raters (for the jury as well as for students).

The "problem" is that in our case, we had an exceptionally large number of ideas (n=310), and in the case of students ratings also an exceptionally large number of raters (k ~ 46) compared to other studies which reported on ICC. It has been shown in previous studies that the inter-rater reliability coefficients are not independent from the number of raters and the number of targets [6]. We will briefly explain this fact for the ICC (1, r) coefficient, which we applied for our rating results. Per definition [23]:

$$\mathrm{ICC}(1,\mathrm{r}) = \frac{\mathrm{BMS\text{-}WMS}}{\mathrm{BMS}} = 1\text{-}\frac{\mathrm{WMS}}{\mathrm{BMS}} \tag{1}$$

where BMS denotes the "between-targets mean square" and WMS the "within-target mean square". Let us assume we have n ideas and k ratings per idea. BMS and WMS are calculated as follows, with x_{ij} being the rating j for idea i:

$$\mathrm{BMS} = \frac{\mathrm{k}}{\mathrm{n\text{-}1}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\overline{\mathrm{x}_\mathrm{i}}\text{-}\overline{\mathrm{x}})^2 \tag{2}$$

$$\mathrm{WMS} = \frac{1}{\mathrm{n\,(k\text{-}1)}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{k}} (\mathrm{x_{ij}}\text{-}\overline{\mathrm{x_i}})^2 \tag{3}$$

We now assume that we double the number of raters, and the new raters perform exactly the same rating as the initial set of raters. With 2k raters we get a new ICC which is definitely higher than the initial ICC:

$$\mathrm{ICC_{new}} = 1\text{-}\frac{\mathrm{k\text{-}1}}{\mathrm{2k\text{-}1}} \frac{\mathrm{WMS_{old}}}{\mathrm{BMS_{old}}} > \mathrm{ICC_{old}} \tag{4}$$

$$\mathrm{ICC_{new}} = \frac{(\mathrm{k\text{-}1})\mathrm{ICC_{old}}+\mathrm{k}}{\mathrm{2k\text{-}1}} \tag{5}$$

The assumption that the new raters perform exactly the same ratings as the initial raters is certainly very unlikely for a small set of raters. On the other hand, our data confirmed that actually almost all 310 idea ratings from students were normally distributed (KS-test, H_0 rejected in only 1 of 310 cases, $\alpha = 0.05$, average p-value 0.633). Now if ratings for each idea are normally distributed, we can well argue that for instance 100 raters would perform very similar ratings as another group of 100 raters. If

we split 46 raters (students rating) into two groups and each group of 23 would theoretically perform an identical rating, ICC for each of the two groups would be, according to equation (5):

$$\mathrm{ICC}_{23} = \frac{45}{22}\mathrm{ICC}_{46}\text{-}\frac{23}{22} \tag{6}$$

An ICC_{46} coefficient of 0.8 would be equivalent to an ICC_{23} coefficient of 0.59, which is below the 0.7 threshold proposed for good reliability. A conclusion of this is that ICC coefficients are likely to be large when we have a large set of raters, as in the case of our students rating.

The other peculiarity is that in our case we have quite a large number of ideas (n=310). Using a similar logic as explained above, we can develop the following reasoning: If we double the number of ideas and the judges rate the new ideas exactly as they rated the initial set of ideas, e.g. because the ideas are very similar, then we can calculate the following:

$$\mathrm{ICC}_{\mathrm{new}} = 1\text{-}\frac{2\mathrm{n}\text{-}1}{2\mathrm{n}\text{-}2}\frac{\mathrm{WMS}_{\mathrm{old}}}{\mathrm{BMS}_{\mathrm{old}}} < \mathrm{ICC}_{\mathrm{old}} \tag{7}$$

According to equation (7), increasing the number of ideas actually has just the opposite effect on the ICC coefficient as increasing the number of raters. We thus suppose that the low inter-rater reliability of our jury evaluations was caused at least partially by (1) the large number of ideas and (2) the small number of raters (k=3).

2.3 Rating Validity

Rating validity is a related but still different concept than reliability: Ratings from a panel of judges may seem reliable (high consensus among raters) but they might apply the wrong target function, i.e. might come to wrong conclusions in terms of which ideas are good or bad. Measuring validity is very difficult if not impossible in innovation contests, because the true value of an idea can hardly be measured before the innovation is implemented and market success can be proven. Therefore, the idea of the CAT is that the *use of experts* and *high reliability of their assessments* should provide a meaningful approximation of the rating validity [1, p. 41ff]. A systematic evaluation process and the application of well-grounded expert knowledge are deemed to be sufficient measures for achieving a valid assessment [14].

Once a jury of experts has performed their evaluation and the reliability of their ratings has shown to be high, one can also assess the validity of evaluations from non-expert raters (e.g. participants): If participants were able to identify the best ideas in conformity with expert judgments, we talk about concurrent validity and we also deem the participants rating as valid [21], [22].

3 A Conceptual Model for Idea Selection

3.1 The value function

In most innovation contests, rating validity cannot be directly measured. At the point of running the contest, the true "value" of the ideas can only be predicted, but we can hardly know whether one or another idea will in fact become a commercial success.

The beauty about models is that we don't have to adhere to all real-world difficulties: We can assume that we know the value of the ideas, and we can then compare the rating of an idea against its true value. And for determining the value of the ideas, we can assume that the value of the ideas can be formulated as a parametric function. In order to construct a realistic value function, let us consider a famous innovation contest, the Netflix Prize challenge [18]: The contest was organized by the online

movie rental company Netflix and was run between October 2006 and September 2009. The task was to develop a new algorithm for their movie recommendation system that was at least 10% better than the former recommendation algorithm [3]. Selecting the best idea was obviously not really a problem because the evaluation criterion was clearly measureable: The quality of the algorithm is determined by calculating the square root mean error between the prediction and the actual movie rating. By comparing the contest proposals with their algorithm they could directly measure the improvement in percent. Figure 2 shows the best 100 proposals (among 44,014 submissions) [18]. We can interpret this as the *value function* of the proposals.

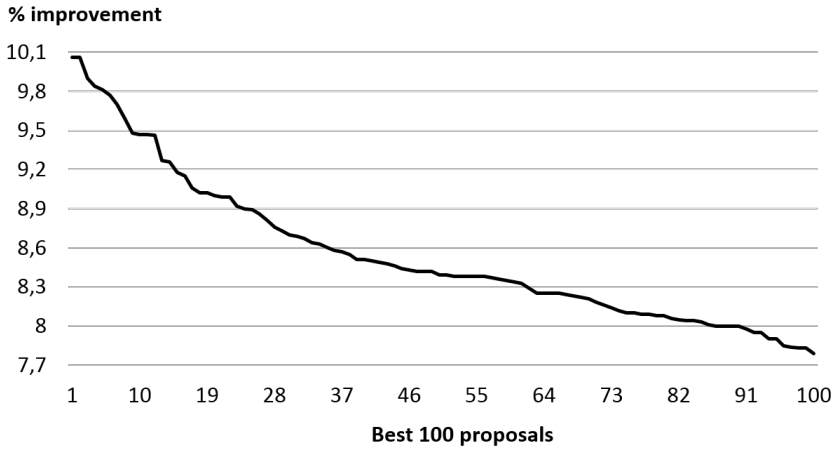

Figure 2. Value function of the Top-100 proposals from the Netflix Prize Challenge [28]

Interestingly, the Netflix Prize challenge is one example where the value of a proposal is directly measureable, and therefore helps to demonstrate a realistic distribution of a possible value function. For our purposes, we will consider three value functions for subsequent modelling and testing. We call them A, B, and C (figure 3). All three functions show the value of 100 ideas from best (left) to worst (right). For simplicity, the best idea has value 1 and the worst idea value 0, i.e. we have normalized the value of ideas.

Function B pretty much looks like the Netflix function for the 100 best ideas. We additionally construct two more functions, one with a steeper decline in the best ideas (function C) and one with a more moderate decline in the best ideas (function A).

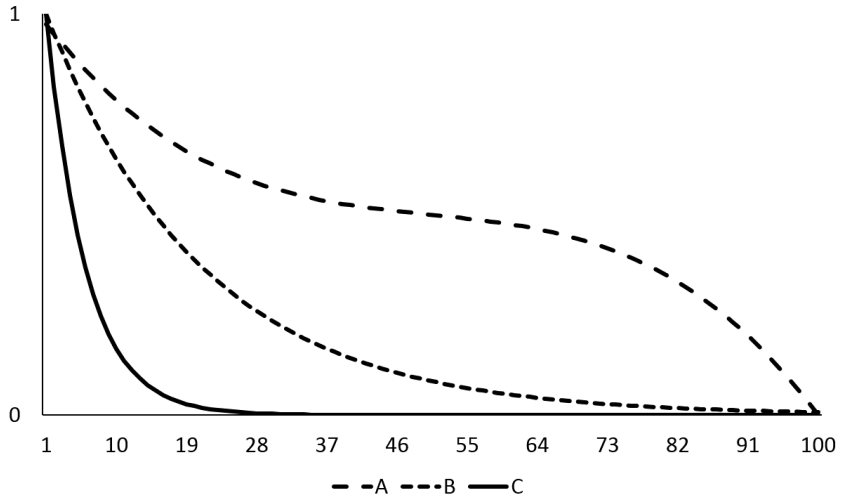

Figure 3. Hypothetical value functions for three innovation contests A, B and C

The functions shown in figure 3 are built as follows (n=100):

$$A(i) = 0{,}6\text{-}0{,}4\left(\frac{2i\text{-}n}{n}\right)^3 \text{-}0{,}2\frac{i}{n} \quad (8)$$

$$B(i) = \exp\left(\frac{5\text{-}5i}{n}\right) \quad (9)$$

$$C(i) = \exp\left(\frac{20\text{-}20i}{n}\right) \quad (10)$$

3.2 Idea Selection

Unlike the Netflix challenge, the usual case in most innovation contests is that the idea selection process does not result in a single chosen idea, but rather a set of the best ideas [8]. A prominent example is the Google 10-to-the-100, where ultimately the best five ideas were selected and granted a total award of 10 million dollars [20].

Now, if we wanted to select the best 10 ideas out of 100, we suppose that idea selection should be easiest in contest C because the value of the best 10 ideas is relatively much higher than the other 90 ideas, as compared to contests A and B. This understanding is based on the following logic: Even if raters are not able to identify the true value of the ideas, they should still be able to determine an approximation which is close to the true value. If the rating criteria are suitable for determining the true value, the raters' evaluations should be distributed normally around the true value [23].

The variance of this normal distribution can be caused by a set of different biases [11]: (1) Raters may not have an identical understanding of the rating criteria, and (2) raters may interpret scales differently, i.e. some raters are generally more positive (trustful) or negative (skeptical) than others. In addition, external effects will influence the ratings because (3) common rating criteria are qualitative measures, i.e. not absolutely measureable but rather depending on a comparison of the individual understanding of the idea compared with past experiences and drawn analogies. The first two biases would be "rater-specific", whereas the third would be called "dyad-specific" as it specific for each rater-idea couple [11].

In our model, ratings x_{ij} for idea i from rater j can be formulated as

$$x_{ij} = \mu_i + r_j + d_{ij} + \varepsilon_i \quad (11)$$

where μ_i is the theoretical true value of idea i, r_j is the rater-specific bias and d_{ij} is the dyad-specific bias for each rater-idea couple. Both the rater-specific bias as well as the dyad-specific bias are assumed to be normally distributed with a mean of zero and a variance of σ_R^2 and respectively σ_D^2 [23]. In addition, there might be an idea-specific error-term ε_i which could be caused by a systematic over- or under-evaluation of ideas (we can imagine that nicely visualized ideas would obtain better ratings than poorly presented ideas, even if they perform equally on all specified evaluation criteria). Because our focus in this study is not on optimizing rating criteria or criteria compliance, we set $\varepsilon_i = 0$.

Using assumptions for the distribution of rater biases (i.e. defining values for σ_R^2 and σ_D^2), we are now able to predict which ideas would be selected in the three contests A, B and C. We can then calculate the value of the selected ideas and compare it with the value of the best ideas. The value difference between the best ideas and the selected ideas would be the "*lost value*" from the selection.

4 Simulations and Findings

Our objective is now to test and visualize the impact of certain design parameters on the selection of the best ideas in our model in order to gain insights for the idea selection problem, and in particular for the idea selection in our contest. For this purpose, we run Monte Carlo simulations with 1,000 iterations for each of the following scenarios: The three value functions A, B and C with $\mu_i \in [0,1]$; number of raters k_1=4, k_2=8 and k_3=20; number of selected best ideas s_1=5, s_2=10 and s_3=25, and a total variance ($\sigma_R + \sigma_D$) ranging from 0.1 to 1.0 in steps of 0.1. We also distinguish between rating consistency (all raters evaluate each idea), inconsistency with fixed allocations (a subset of raters evaluate the same subset of ideas) and with random allocation (a subset of raters evaluate a randomly assigned subset of ideas). Overall, the combination of all parameters resulted in 810 different scenarios. In order to refrain from more complexity, we kept the number of ideas fixed at n=100 and we simply divided the total variance into equal halves ($\sigma_R^2 = \sigma_D^2 = \sigma^2/2$). In our contest, the rater-specific bias from students was found to make ~40% of their overall variance of ratings per idea, i.e. using $\sigma^2/2$ in our model does not seem to be unrealistic.

For each of the 1,000 iterations in each of the 810 scenarios we calculated two values: The lost value and the ICC coefficient. In contrast to the ICC which shows the reliability of the ratings, the lost value is actually a measure for the validity of the rating. Based on the value functions which we constructed, we were able to compute the degree of validity for each rating scenario. The higher the columns (figures 4 and 5), the lower the validity of the rating, expressed in the overall lost value from the idea selection decision. Results discussed in the following are average values from all 1,000 iterations. The x-axis shows the total bias variance from $\sigma^2 = 0.1$ to $\sigma^2 = 1.0$.

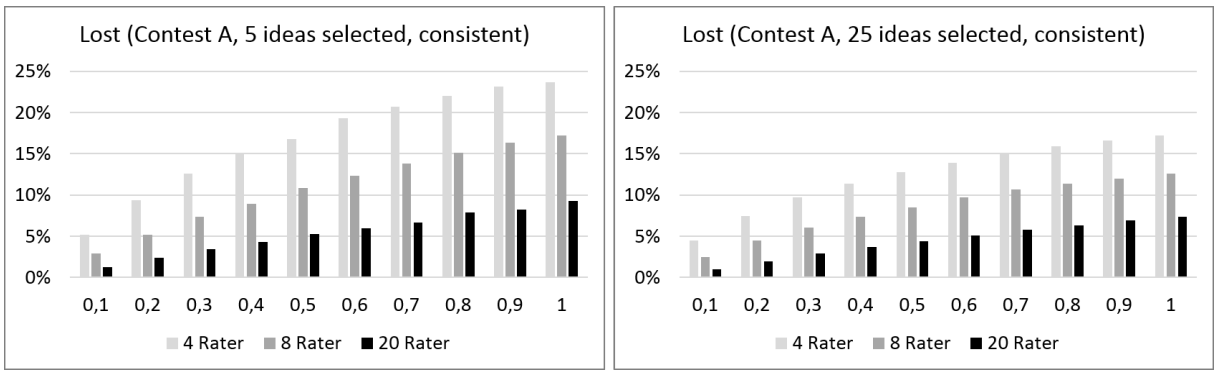

Figure 4. Lost value (by idea selection): Contest A

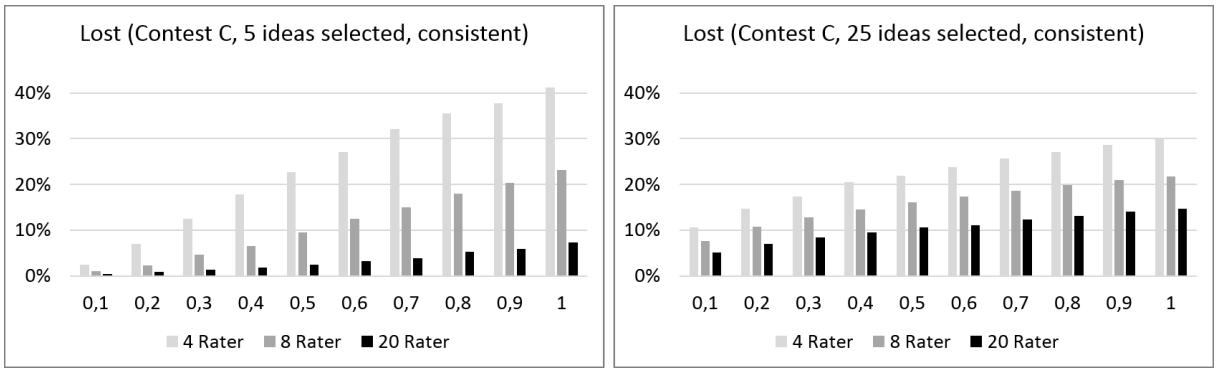

Figure 5. Lost value (by idea selection). Contest C

From all scenarios which we computed, figures 4 and 5 illustrate the results which are the most remarkable. Our calculations showed that the "in-between" scenarios (contest B in between A and C; s=10 selected ideas in between 5 and 25) produce results which are in between those shown in figures 4 and 5. For interpreting our findings we therefore concentrate on the results which most clearly exhibit the differences between the scenarios.

At first sight, the results confirm two aspects which we had expected: (1) The larger the rater bias, the higher the overall lost value, and (2) more raters are better than few at determining the best ideas, at least for the same rater bias. A bit surprising at first sight is that selecting the best ideas in contest C was obviously not "easier" than in contest A. The lost value in contest C is higher than in contest A in most scenarios (only when we select 5 ideas, the lost value in contest C is lower than in contest A, at least for some σ^2 values). We can explain this observation by the fact that the calculation is done as

$$v_{lost} = (\sum v_{best} - \sum v_{selected}) / \sum v_{best} \quad (12)$$

and the effect of making "wrong" decisions is smaller for contest A than for contest C. Hence, although in contest C we should be able to better identify the best ideas, the larger effect from making wrong decisions makes the idea selection more difficult in the end.

Another interesting observation can be made by comparing the *maximum level of variance* which is *acceptable* if we want to achieve a *certain degree of rating validity*. In all scenarios, 20 raters with a bias variance of even 1.0 performed better at idea selection than 4 raters with a bias variance of 0.3. In other words, a community of 20 raters which has a bias 3 times as large as a jury of 4 raters would still be expected to perform better at selecting the best ideas, at least for the applied value functions.

Looking at the reliability of the ratings, it is apparent that we did not achieve ICC coefficients > 0.7 in any of the 4-rater scenarios – not even for the smallest variance of 0.1 (see figure 6). This result confirms our discussion from section 2.2 where we argued that ICC coefficients for a small number of raters and contests with many ideas (in this case 100) would be low in most cases. It also explains why we got such low ICC coefficients for the jury rating in our contest, where we actually had triple the number of ideas (n=310) and only 3 raters per idea.

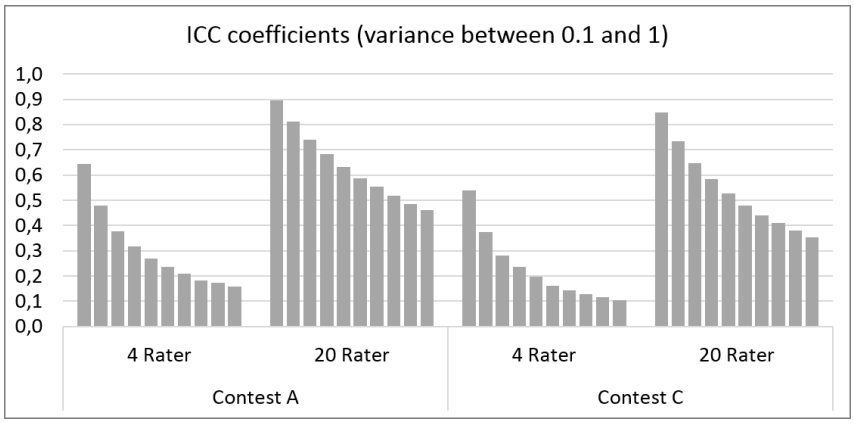

Figure 6. ICC (2,r) coefficients for contests A and C, comparison of 4 and 20 raters

By comparing figure 6 with figures 4 and 5 we observe a high correlation between rating performance and the inter-rater reliability (Pearson correlation test confirmed significance at p=0.01 level). But this correlation cannot be interpreted as a linear relationship: 4 raters in contest A had a lost value of ~10% at $\sigma^2 = 0.3$ while 20 raters in contest C had a lost value of ~10% at $\sigma^2 = 0.4$. The corresponding ICC coefficients are 0.38 and 0.58, i.e. the ICC coefficients for almost the same rating performance are found to be quite different.

Comparing consistent ratings (raters evaluate each idea) with inconsistent ratings (raters evaluate a subset of ideas) offers additional insights: First of all, consistent rating performed much better than both types of inconsistent ratings. For instance, a consistent rating with $\sigma^2 = 1$ resulted in a similar rating performance as inconsistent ratings with $\sigma^2 = 0.6$ (see figure 7). Comparing both types of inconsistent ratings, i.e. fixed vs. random allocation of ideas, we could not find a significant difference.

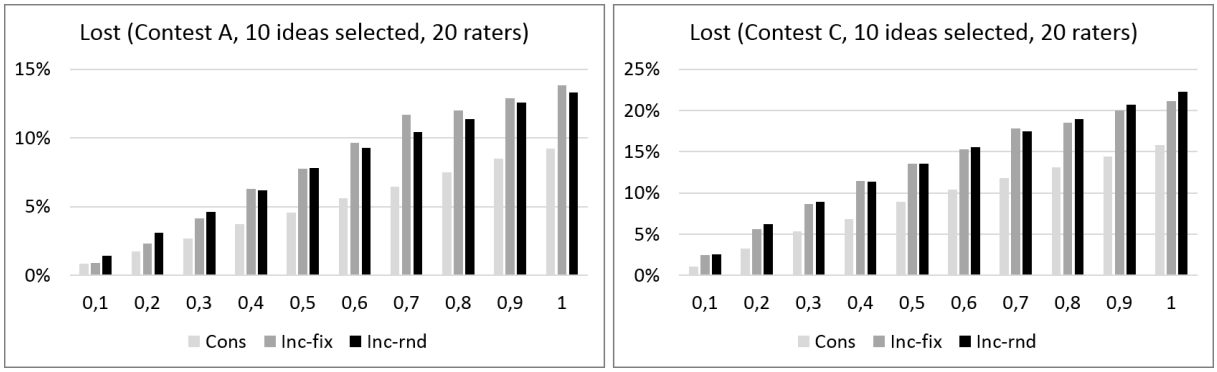

Figure 7. Lost value compared for consistent, inconsistent fixed and random

We also compared the ICC coefficients for the same scenarios (figure 8). Especially for contest C it is interesting that the ICC is lowest for inconsistent ratings with fixed idea allocation, although fixed allocation has performed slightly better in terms of value selection.

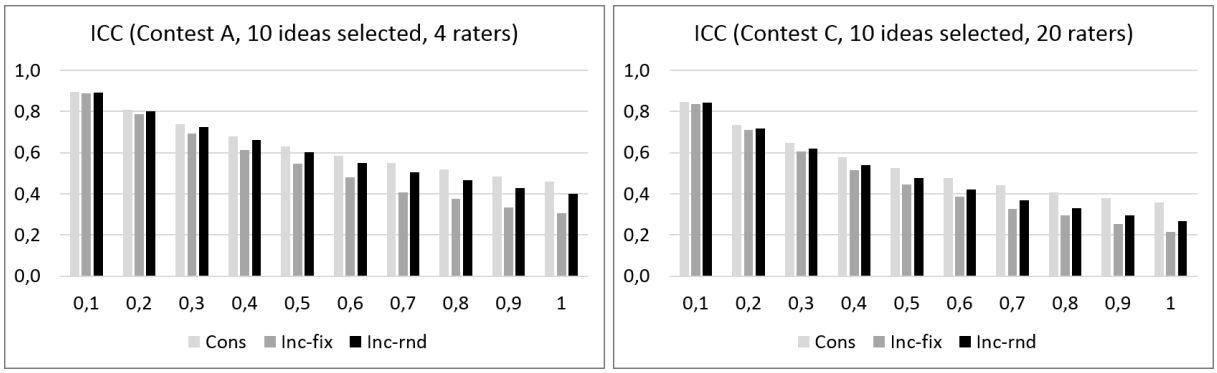

Figure 8. ICC coefficients compared for consistent (ICC (2, r)), inconsistent fixed and random (ICC (1, r))

5 Discussion & Conclusion

Although from our modelling results we could not gain renewed confidence with the rating results from our contest, the presented findings reveal various insights which could be considered for subsequent innovation contests – in a sequel of the contest at our university, and also in other large-scale innovation contests. First of all, the results have proven that rating reliability is a very useful proxy for the validity of a rating as the computed ICC coefficients and rating performance in our model are highly correlated. Nevertheless, the ICC coefficients should be interpreted with caution: Commonly used thresholds seem to be adequate when a jury of experts evaluates a small number of ideas. But once we have a larger number of ideas, ICC coefficients will most probably be low due to the difficulty to clearly distinguish between the perceived value of ideas. Unfortunately, we cannot really offer a solution to this problem: Even though from our simulations we can interrelate ICC values from 810 different idea contests with their corresponding rating validity (or lost value), the set of assumptions used for our model is still rather limited, compared to the potential configuration space of innovation contests in practice. In addition, we have seen from our modelling results that there is clearly no linear relationship between ICC coefficients and rating validity, i.e. a systematic correlation between model parameters, ICC coefficients, and rating validity seems to be a complex endeavor. Developing new ICC thresholds for large-scale idea contests with many raters would be highly desirable but would also require a significant extension of non-static model parameters.

Regardless the above mentioned limitation, we can offer strong support for engaging a larger community of raters. While this statement can be qualitatively explained by the law of large numbers, our simulations go one step further and provide numeric insights: For *all scenarios* – and the value functions A, B and C were actually quite different – we found that a community of 20 raters would outperform a jury of 4 raters, even if the rating bias of the community was *three times* higher. This finding reassures the relevance of "open evaluation" approaches [8], especially because it is obviously much easier to recruit an extra couple of sufficiently skilled and motivated raters, compared to taking measures for significantly improving the rating ability of your raters at hand.

Our simulations confirmed another qualitative statement which relates to the set-up of idea selection in innovation contests: Ideally you would use a large group of expert raters who *consistently* evaluate the entire set of ideas. Consistency means that each rater evaluates all ideas. From figure 8 we can get an impression of the advantages of consistent ratings with regards to rating reliability. What is interesting now is that the rather small advantage of consistent ratings expressed in ICC values turned out to be much larger when looking at the rating validity (see figure 7): From all scenarios, consistent ratings performed *on average 40% better* at value selection compared to the inconsistent ratings. The problem

is that a request for rating consistency will hardly be complied with, once we talk about innovation contests with very large numbers of ideas. For instance in the Google 10-to-the-100 contest, a team of 3,000 employees was engaged to evaluate an overwhelming amount of 150,000 ideas – a task which would have never been accomplished with a consistent rating [15].

We finally want to stress that rating performance clearly depends on the rater bias, and logically the smaller the bias the better the results. In our model we simply used σ^2 as the variance of the overall rating bias and split this into equal parts for the rater-specific and the dyad-specific bias. On this aspect we would call for additional research which could investigate measures to improve the alignment of raters (i.e. reducing the variance of ratings for the same idea). Such measures seem to bear great potential for improving the idea selection process, and it would be particularly useful for community ratings where we usually have rather little control on the rating abilities of "non-expert" raters.

Coming back to the idea rating results in our innovation contest with students, it became clear that the difference between ICC values from the jury and the student evaluation can mainly be explained by the large number of student raters and the small number of "expert" raters per idea. Still, the results cannot be deemed satisfactory. The first question is whether the jury was sufficiently skilled to realistically evaluate the ideas. With regards to evaluating ideas for *digital city services*, our 27 research assistants may have had not enough common understanding about how to apply the rating criteria. A briefing workshop and maybe a joint evaluation trial-run could have been useful. The second question is whether the jury members were sufficiently motivated to concentrate on their task to evaluate the ideas. Expert rater motivation is a general problem, because we can imagine that highly skilled people with great expertise are usually very busy and have very limited time for their evaluation task. The last question is whether the students were actually able to assess the ideas on all defined criteria. Especially technical and economic feasibility are criteria which they may not be sufficiently knowledgeable about. Here, an idea for the future would be to limit the community evaluation to the more straightforward criteria, derive a short-list of ideas, and then have the experts evaluate the short-listed ideas on the more difficult criteria, or on a wider set of criteria.

6 References

[1] Amabile, TM (1996): Creativity in context: Update to the social psychology of creativity. Westview Press.

[2] Blohm, I.; Bretschneider, U; Leimeister, JM; Krcmar, H (2011): Does collaboration among participants lead to better ideas in IT-based idea competitions? An empirical investigation. Int. J. of Networking and Virtual Organisations 2(9):106-122.

[3] Boudreau, KJ; Lacetera, N; Lakhani, KR (2011): Incentives and problem uncertainty in innovation contests: An empirical analysis. Management Science 57(5):843-863.

[4] Bretschneider, U (2012): Die Ideen Community zur Integration von Kunden in die frühen Phasen des Innovationsprozesses. Gabler, Wiesbaden.

[5] Bullinger, AC; Neyer AK; Rass, M; Moeslein, KM (2010): Community-Based Innovation Contests: Where Competition Meets Cooperation. Journal of Creativity and Innovation Management 19(3):290-303.

[6] Cortina, JM (1993): What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 78(1):98-104.

[7] Ebner, W; Leimeister, JM; Krcmar, W. (2009): Community engineering for innovations: The ideas competition as a method to nurture a virtual community for innovations. R&D Management 39(4):342-356.

[8] Haller, J. (2013): Open Evaluation. Integrating users into the selection of new product ideas. Gabler, Wiesbaden.

[9] Hansen, EG; Bullinger, AC; Reichwald R (2011): Sustainability innovation contests: Evaluating contributions with an eco impact-innovativeness typology. Int. J. Innovation and Sustainable Development 5(2/3):221-245.

[10] Hartmann, M; Bretschneider, U; Leimeister JM (2013): Patients as innovators. The development of innovative ideas with the Ideenschmiede. *11th Int. Conference on Wirtschaftsinformatik*

[11] Hoyt, WT (2000): Rater bias in psychological research. When is it a problem and what can we do about it? Psychological Methods 5(1):64-86.

[12] Innovation Contest Inventory: http://innovationresearch.de, accessed 09-15-2013.

[13] Jeppesen, LB; Lakhani KR (2010): Marginality and problem-solving effectiveness in broadcast search. Organization Science 21(5): 1016-1033.

[14] Kaufman JC; Baer, J; Cole, JC; Sexton JD (2008): A comparison of expert and non-expert raters using the consensual assessment technique. Creativity Research Journal 20(2):171-178.

[15] Klein, M (2012): Enabling large-scale deliberation using attention-mediation metrics. Computer Supported Cooperative Work 21(4/5):449-473.

[16] Kristensson, P; Gustafsson, A; Archer, T (2004): Harnessing the creative potential among users. Journal of Product Innovation Management 21(1):4-14.

[17] Leimeister, JM; Huber, M; Bretschneider, U; Krcmar, H. (2009): Leveraging crowdsourcing: Activation-supporting components for IT-based ideas competition. Journal of Management Information Systems 26(1):197-224.

[18] Netflix Prize: www.netflixprize.com, accessed 09-15-2013.

[19] Piller, FT; Walcher, D (2006): Toolkits for idea competitions: A novel method to integrate users in new product development. R&D Management 36(3):307-318.

[20] Project 10-to-the-100: www.google.com/campaigns/project10tothe100, accessed 09-15-2013.

[21] Reinig, BA; Briggs, RO; Nunamaker, JF (2007): On the Measurement of Ideation Quality. Journal of Management Information Systems 23(4):143-161.

[22] Riedl, C; Blohm, I; Leimeister, JM; Krcmar, H (2010): Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right. *ICIS 2010 Proceedings.*

[23] Shrout, PE; Fleiss, JL (1979): Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86(2):420-428.

[24] Walcher, D (2007): Der Ideenwettbewerb als Methode der aktiven Kundenintegration: Theorie, empirische Analyse und Implikationen für den Innovationsprozess. Gabler, Wiesbaden.

[25] Wenger, JE (2012): Gewinngestaltung bei Innovationswettbewerben. Theoretische und praktische Betrachtung. Gabler, Wiesbaden