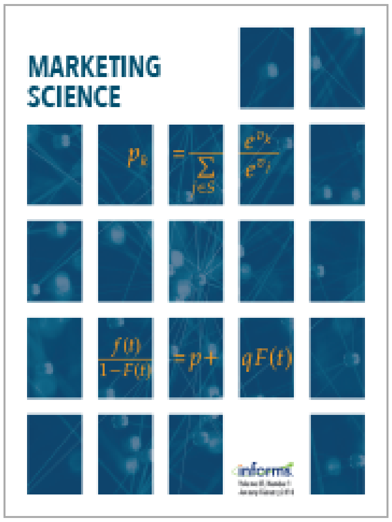


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Adaptive Idea Screening Using Consumers

Olivier Toubia, Laurent Florès,

To cite this article:
Olivier Toubia, Laurent Florès, (2007) Adaptive Idea Screening Using Consumers. Marketing Science 26(3):342-360. https://doi.org/10.1287/mksc.1070.0273

MARKETING SCIENCE
Vol. 26, No. 3, May–June 2007, pp. 342–360
ISSN 0732-2399 | EISSN 1526-548X | 07 | 2603 | 0342

DOI 10.1287/mksc.1070.0273

Adaptive Idea Screening Using Consumers

Olivier Toubia
Columbia Business School, Uris Hall, Room 522, 3022 Broadway, New York, New York 10027-6902,
ot2107@columbia.edu

Laurent Florès
Laboratoire INSEEC and CRMMETRIX, 700 Plaza Drive, 2nd Floor,
Secaucus, New Jersey 07094, lflores@crmmetrix.com

Following a successful idea generation exercise, a company might easily be left with hundreds of ideas generated by experts, employees, or consumers. The next step is to screen these ideas and identify those with the highest potential. In this paper we propose a practical approach to involving consumers in idea screening.

Although the number of ideas may potentially be very large, it would be unreasonable to ask each consumer to evaluate more than a few ideas. This raises the challenge of efficiently selecting the ideas to be evaluated by each consumer. We describe several idea-screening algorithms that perform this selection adaptively based on the evaluations made by previous consumers. We use simulations to compare and analyze the performance of the algorithms as well as to understand their behavior. The best-performing algorithm focuses on the ideas that are the most likely to have been misclassified (as "top" or "bottom" ideas) based on the previous evaluations, and avoids discarding ideas too fast by adding random perturbations to the misclassification probabilities. We demonstrate the convergent validity of this algorithm using a field experiment, which also confirms the convergence pattern predicted by simulations.

Key words: innovation; marketing research; marketing surveys; marketing tools; new product research; product development
History: This paper was received November 15, 2005, and was with the authors 5 months for 2 revisions; processed by Arvind Rangaswamy.

1. Introduction

Idea generation is critical to new product development. It belongs to the "fuzzy front end" of the development process, recognized as a key leverage point for a firm (Dahan and Hauser 2001a, Hauser et al. 2006). A variety of idea generation methods have been introduced since the 1950s. The most popular is probably brainstorming (Osborn 1957). Other traditional examples include lateral thinking (De Bono 1970), synectics (Prince 1970, Gordon 1969), and six thinking hats (De Bono 1985). Recent developments include electronic brainstorming (Nunamaker et al. 1987; Gallupe et al. 1991, 1992; Dennis and Valacich 1993; Valacich et al. 1994), ideation templates (Goldenberg et al. 1999a, b; Goldenberg and Mazursky 2002), and incentives-based idea generation (Toubia 2006). With the development of Internet-based tools, companies are increasingly involving their own consumers in idea generation (*Forbes* 2005).

Depending on the method used, a successful idea generation exercise may result in up to hundreds of ideas generated by experts, consumers, or employees. The number of ideas appears even more likely to be large when consumers are involved in the process.[1] The new product development team is then left with the daunting task of screening these ideas in order to focus its limited resources on those with the highest potential. The selected ideas will be refined and translated into specific features or integrated products. In our field experiment, examples of consumer-generated ideas on how to improve cellular phones (see Table 2) included "There would be a way to ftp data files," "Download movies with your phone and project them on the wall so it seems like you're at the theater," etc. One traditional approach to idea screening is to ask one or a few experts to go over the transcripts of ideas and evaluate them (Urban and Hauser 1993). However, experts' judgments might not always reflect consumers' needs and preferences.[2]

In this paper we propose a practical approach to involving consumers in idea screening. Although the number of ideas to be screened may potentially be very large (especially if a large number of consumers have been involved in the idea generation process), it would be unreasonable to ask each consumer to evaluate more than a few ideas, especially if the evaluations are to be performed online in a noncontrolled

[1] An extreme example is Staples' recent organization of a competition among consumers to generate new product ideas. Eighty-three hundred ideas were submitted (*The Economist* 2005).

[2] Indeed, decisions in the subsequent stages of the development process are often supported by marketing research tools such as conjoint analysis, focus groups, or pretest market forecasting (Urban and Hauser 1993).

environment (Dahan and Hauser 2001b). This raises the challenge of efficiently selecting the ideas to be evaluated by each consumer in order to converge to the best ideas as quickly (i.e., with only few respondents) and reliably as possible. We assume that the evaluations are done online and sequentially, allowing the selection to be performed adaptively based on the evaluations made by previous consumers.

We propose and explore several algorithms for adaptive idea screening. We assume that each idea appeals to an unknown proportion of consumers. Our estimate of this proportion follows a beta distribution with parameters depending on the previous evaluations. We assume that the team's objective is to identify the top m ideas out of a given set. We use simulations to compare and analyze the performance of the algorithms, as well as to understand their behavior and the drivers of differences in performance. We demonstrate the convergent validity of the best-performing algorithm using a field experiment, which also confirms the convergence pattern predicted by simulations. Note that our field experiment focuses on convergence and convergent validity, and that we rely on simulations to compare the performance of the different algorithms.

A problem with some similarities to ours was studied in the educational testing literature by Bradlow and Wainer (1998). Bradlow and Wainer consider subjective tests (e.g., essays) raters by human judges (on a continuous scale), resulting in binary pass/fail decisions (such that only candidates with an average grade above a predefined cutoff pass). They consider a situation in which the rescoring of some tests is possible after all tests have been rated by a fixed number of judges and initial pass/fail decisions have been made, and study the problem of allocating judges in the rescoring phase (e.g., which essays should be graded again). They find, using a modeling setup different from ours,[3] that for tests in which the number of initial failers and passers are approximately equal, a reasonable strategy is to rescore only examinees near the cutoff score (they compare this strategy to one where only failures are rescored).

Beyond the differences in modeling approach, context, and type of evaluations, two fundamental differences between our problem and the one studied by Bradlow and Wainer are that (1) the number of previous evaluations per item is constant across items in the latter (same initial number of raters on each essay) and different in the former (different number of previous evaluations per idea) and (2) allocation decisions are made once in the latter versus many times (once for each consumer) in the former. Given these differences, Bradlow and Wainer's work is not directly applicable to our problem. However, we will use it as an initial building block for some of our algorithms.

This paper is structured as follows. We introduce the idea selection algorithms in §2. In §3 we report the results of a series of simulations designed to study the performance and behavior of these algorithms. We report the results of our field experiment in §4. We describe a managerial application of our research in §5 and conclude in §6.

2. Algorithms for Adaptive Idea Selection

Notations and Definitions

As mentioned earlier, we assume that our goal is to select a fixed number of ideas to be brought to the next stage of the new product development process, i.e., to identify the top m ideas out of a set of I previously generated ideas. In order to achieve this goal, we ask different consumers to evaluate different subsets of k ideas. For simplicity, we assume that consumers provide binary evaluations of the ideas, i.e., they indicate which ideas they believe to be "good." Note that we do not restrict the definition of a "good" idea. It can be specified by the researcher and should be explicitly given to the consumers before they start their evaluations. Note also that we show in Appendix A how our framework could be extended to nonbinary evaluations. We leave to future research the extension to other screening goals, such as identifying all ideas above a predefined threshold.

Let us define $(p_i)_{i\in\{1,\dots,I\}}$ as the probability that a randomly selected consumer will classify idea i as a good idea. We use p_i as a measure of the quality of the idea, i.e., our goal is to identify the m ideas with the highest probabilities.

Let us define the following:

$(n_{Si})_{i\in\{1,\dots,I\}}$ = number of respondents who have evaluated idea i and classified it as a good idea. (S stands for "success.")

$(n_{Fi})_{i\in\{1,\dots,I\}}$ = number of respondents who have evaluated idea i and did not classify it as a good idea. (F stands for "failure.")

n_{S0}, n_{F0}: parameters of our prior on p_i, assumed to follow a beta distribution $Beta(n_{S0}, n_{F0})$.

$(\hat{p}_i)_{i\in\{1,\dots,I\}}$ = our estimate of p_i, based on the previous evaluations. Given our beta prior and the fact that the evaluations follow a binomial likelihood, the posterior on p_i follows another beta distribution: $Beta(n_{S0}+n_{Si}, n_{F0}+n_{Fi})$.[4] Our point estimate of p_i is simply the expected value of this distribution: $\hat{p}_i = (n_{Si}+n_{S0})/(n_{Fi}+n_{Si}+n_{S0}+n_{F0})$.

[3] The tth evaluation y_{ijt} of examinee i by rater j being on a continuous scale, Bradlow and Wainer assume that $y_{ijt}=\mu+\alpha_i+\beta_j+\varepsilon_{ijt}$ where $\alpha_i \sim \mathrm{N}(0,\sigma_\alpha^2)$, $\beta_j \sim \mathrm{N}(0,\sigma_\beta^2)$, and $\varepsilon_{ijt} \sim \mathrm{N}(0,\sigma^2)$. An examinee passes the test if $\bar{y}_{ij} \geq c$ where c is a predefined cutoff.

[4] Beta priors are conjugates for binomial likelihoods (Gelman et al. 1995).

Note that our definition of quality ignores other important criteria such as cost, feasibility, or fit with the company's core competencies (Ozer 2005). This can be addressed by subjecting the subset of ideas selected using our approach to a second round of evaluations based on these other criteria (likely to be performed by a small set of experts in a traditional fashion). Consumer acceptance being a necessary condition for success, the initial screening performed by consumers would probably make this second round of evaluations easier with only a reduced risk of leaving out potentially fruitful ideas.

Note finally that this framework may be extended to account for the existence of noise in the evaluations. In the simplest case, the amount of noise may be assumed to be constant across ideas and across consumers. For example, if a consumer produces a random evaluation (positive with probability 0.5) with probability 0.2, the observed probability for idea i is $0.8 * p_i + 0.2 * 0.5$, where p_i is the "true" probability. Such monotonic transformation would not change the identity of the top m ideas; however, it would make their identification harder by reducing the amount of variation across ideas. We leave the extension to cases in which the amount of noise is assumed to vary across ideas and/or across consumers to future research.

We assume that the sets of ideas presented to the first $(n-1)$ consumers, as well as the corresponding evaluations, are available when selecting the subset of ideas to be presented to the nth consumer.[5] This is the case if the evaluations are done online (as in our field experiment). Consumers could be invited to participate by e-mail, or directed to the evaluation site from the company's main site.

Objective Function and Performance Metrics

The problem of identifying the top m ideas can be viewed as that of correctly classifying ideas into two groups: the group composed of the m ideas with the highest associated probabilities, and its complement. We will refer to these two groups as the top and bottom groups, respectively. We index the m ideas with the highest probabilities as $i_1, \ldots, i_m$ and the others as $i_{m+1}, \ldots, i_I$ such that $\hat{p}_{i_1} \geq \cdots \hat{p}_{i_m} \geq \hat{p}_{i_{m+1}} \geq \cdots \hat{p}_{i_I}$.

We consider two performance metrics:

1. A *hit rate*, defined as the number of ideas estimated to be in the top m that are correctly classified. Although it is easy to interpret, this metric is discrete, and hence does not take into account the actual quality of the ideas.

2. The average true probability (the true probabilities are known in simulations) of the estimated top m ideas: $1/m \sum_{i\in\{i_1,\ldots,i_m\}} p_i$.

Maximizing performance on either of these metrics by adaptively asking N consumers to evaluate k ideas each is a dynamic program with the metric as the objective function, the number of previous positive and negative evaluations for each idea $\{(n_{Si})_{i\in\{1,\ldots,I\}}, (n_{Fi})_{i\in\{1,\ldots,I\}}\}$ as the state, the ideas to be evaluated by the next consumer as decision variables, and the transition between states being given by our current estimates $(\hat{p}_i)_{i\in\{1,\ldots,I\}}$. Unfortunately, the size of the state space is such that the identification of the optimal strategy would be intractable, at least with today's computers and using traditional dynamic programming techniques (Bertsekas 1995). Hence, some approximations are necessary. We first consider the myopic approximations of the dynamic programs corresponding to each of the two performance metrics. Next, we consider additional heuristics, three of which are related to the previous work of Bradlow and Wainer (1998) reviewed in the introduction. A summary of all the algorithms is provided in Table 1.

Common Structure of the Algorithms

All of the algorithms studied in this paper share a common structure. In particular, the following steps are performed in order to select the ideas presented to the next consumer:

Step 1. Estimate the probability associated with each idea: $\hat{p}_i = (n_{Si}+n_{S0})/(n_{Fi}+n_{Si}+n_{S0}+n_{F0})$.

Step 2. Assign a score to each idea, s_i.

Step 3. Select the k ideas with the highest scores.

The algorithms differ only in the method used to compute the scores in step 2. Hence, they all implicitly assume in Step 3 that the improvement in performance achieved by presenting a set of k ideas is monotonically increasing in the sum of the improvements achieved by presenting each of the k ideas independently. In other words, instead of directly assigning a score to each $\binom{I}{k}$ subset of k ideas, they assign one score to each idea and implicitly assign a

Table 1 Summary of the Algorithms

Name of the algorithm	Main characteristic
Myopic Hit Rate	Myopically maximizes the hit rate (number of estimated top m ideas in true top m).
Myopic Average	Myopically maximizes the average true probability of the estimated top m ideas.
Closest to Threshold	Selects the ideas with the estimated probabilities closest to the threshold between the top and bottom groups.
Misclassification Minimization	Selects the ideas most likely to have been misclassified as top or bottom.
Misclassification Minimization with Random Perturbations	Similar to misclassification minimization with the addition of random perturbations to the scores assigned to each idea.
Maximize Right Tail	Selects the ideas most likely to have a probability higher than a predefined threshold.
Random	Selects ideas randomly.

[5] We only consider adaptation *across* consumers. This allows all the ideas presented to a given consumer to be displayed simultaneously.

score to each subset equal to the sum of the scores of its elements. We use such approximation because of the large number of subsets. For example, with $I = 100$ and $k = 10$, there exist $100!/90! \cdot 10! = 1.73 \cdot 10^{13}$ possible subsets of ideas. Future research may investigate heuristics that do not require the enumeration of all subsets. For example, borrowing techniques from the experimental design literature (Kuhfeld et al. 1994, Federov 1972, Cook and Nachtsheim 1980), it may be possible to start with the set of ideas with the highest individual scores and consider replacing each idea with an idea not currently in the set. Such an operation could be repeated until no further improvement is possible, leading to a locally optimal set of ideas.

Myopic Approximations

1. *Myopic Maximization of the Hit Rate*. Our first algorithm myopically maximizes hit rates. The expected hit rate (number of ideas in estimated top m that are in the true top m) is given as a function of the current state $\{(n_{Si})_{i\in\{1,\ldots,I\}}, (n_{Fi})_{i\in\{1,\ldots,I\}}\}$ as follows:

$$\begin{aligned} &H(n_{S1},\ldots,n_{SI},n_{F1},\ldots,n_{FI})\\ &\quad=\int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\ldots,i_m\}}\Bigg(\prod_{j\in\{1,\ldots,I\}\setminus\{i_1,\ldots,i_m\}}1(p_i\geq p_j)\Bigg)\\ &\qquad\cdot\beta_{n_{S0}+n_{S1},\,n_{S0}+n_{F1}}(p_1)\cdots\beta_{n_{S0}+n_{SI},\,n_{S0}+n_{FI}}(p_I)\,dp_1\cdots dp_I, \end{aligned}$$

where 1( ) is the indicator function and $\beta_{n_{S0}+n_{Si},\,n_{S0}+n_{Fi}}$ is the probability density function of $Beta(n_{S0}+n_{Sj},\, n_{S0}+n_{Fj})$. The score assigned to each idea is equal to the expected hit rate that would result from obtaining an additional evaluation on that idea:

$$\begin{aligned} s_i=\hat{p}_i\cdot H\big(&n_{S1},\ldots,n_{S(i-1)},n_{Si}+1,n_{S(i+1)},\ldots,n_{SI},\\ &n_{F1},\ldots,n_{FI}\big)\\ +(1-\hat{p}_i)H\big(&n_{S1},\ldots,n_{SI},n_{F1},\ldots,n_{F(i-1)},n_{Fi}+1,\\ &n_{F(i+1)},\ldots,n_{FI}\big) \end{aligned}$$

In our simulations, we estimated the above integral numerically using 1,000 random draws. Because it requires numerical integration, this algorithm is the least practical, and by far the slowest, of those considered in this paper (all other algorithms can be implemented without noticeable delays between judges).

2. *Myopic Maximization of the Average True Probability of the Estimated Top m Ideas*. Our second algorithm myopically maximizes the second performance metric, i.e., the average probability of the estimated top m ideas. The expected value of this objective function is given as a function of the current state $\{(n_{Si})_{i\in\{1,\ldots,I\}}, (n_{Fi})_{i\in\{1,\ldots,I\}}\}$ as follows:

$$\begin{aligned} &A(n_{S1},\ldots,n_{SI},n_{F1},\ldots,n_{FI})\\ &\quad=\frac{1}{m}\sum_{i\in\{i_1,\ldots,i_m\}}\hat{p}_i=\frac{1}{m}\sum_{i\in\{i_1,\ldots,i_m\}}\frac{n_{S0}+n_{Si}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}. \end{aligned}$$

The score assigned to each idea is:

$$\begin{aligned} s_i=\hat{p}_i\cdot A\big(&n_{S1},\ldots,n_{S(i-1)},n_{Si}+1,\\ &n_{S(i+1)},\ldots,n_{SI},n_{F1},\ldots,n_{FI}\big)\\ +(1-\hat{p}_i)\cdot A\big(&n_{S1},\ldots,n_{SI},n_{F1},\ldots,n_{F(i-1)},\\ &n_{Fi}+1,n_{F(i+1)},\ldots,n_{FI}\big). \end{aligned}$$

Common Characteristic of the Myopic Approximations

We show the following proposition in Appendix B:

Proposition. *Consider the set of ideas classified as top ideas that would remain classified as top after one positive or negative evaluation, and the set of ideas classified as bottom ideas that would remain classified as bottom after one positive or negative evaluation*:

$$\begin{aligned} &\left\{i\in\{i_1,\ldots,i_m\},\ \frac{n_{S_0}+n_{S_i}}{n_{S_0}+n_{S_i}+n_{F_0}+n_{F_i}+1}\geq\hat{p}_{i_{m+1}}\right\}\\ &\qquad\cup\left\{i\notin\{i_1,\ldots,i_m\},\ \frac{n_{S_0}+n_{S_i}+1}{n_{S_0}+n_{S_i}+n_{F_0}+n_{F_i}+1}\leq\hat{p}_{i_m}\right\} \end{aligned}$$

(*we assume, without loss of generality, that* $\hat{p}_{i_1}\geq\cdots\hat{p}_{i_m}\geq\hat{p}_{i_{m+1}}\geq\cdots\hat{p}_{i_I}$).

Each myopic approximation assigns the same score to all ideas in this set.

This proposition implies that each myopic approximation assigns a unique score only to ideas whose classification may change after only one evaluation and assigns the same score s_0 to all other ideas. When fewer than k ideas have a score higher than s_0, these algorithms randomly select the remaining ones from the set with a score of s_0. As a result, we will see in the next section that they do not behave very differently from the random benchmark.

Heuristic Approaches

3. *Closest to Threshold.* Our third algorithm directly and naively applies Bradlow and Wainer's (1998) recommendation ("select ideas near the cutoff"). The score assigned to each idea is equal to the opposite of the distance between its estimated probability and the estimated probability of the closest idea in the other group, i.e.:

- If i is in the "top" group, $s_i=-|\hat{p}_i-\hat{p}_{i_{m+1}}|$.
- If i is in the "bottom" group, $s_i=-|\hat{p}_i-\hat{p}_{i_m}|$,

where $\hat{p}_{i_{m+1}}$ and $\hat{p}_{i_m}$ are, respectively, the highest probability estimate among the current bottom ideas and the smallest probability estimate among the current top ideas.

4. *Misclassification Minimization.* As mentioned earlier, Bradlow and Wainer (1998) study a situation in which the same number of judges is assigned to all examinees in the first step. A more general interpretation of their recommended strategy is that the

examinees that should be rescored are those who are the most likely to have been misclassified in the first round.[6] In their context, this is similar to a cutoff strategy because the number of evaluations on each examinee is the same in the first step. In contrast, we consider a multistage process in which the numbers of evaluations per idea differ. As a result, the probability that an idea has been misclassified is not only driven by the distance to the cutoff, but also by the variance of the corresponding posterior distribution, which is driven by the number of previous evaluations.

Our next heuristic, which we label "Misclassification Minimization," assigns a score to each idea equal to the probability that it has been misclassified based on the previous evaluations. We approximate this probability as follows:[7]

- If i is in the top group, $s_i = \text{Prob}(p_i \leq \hat{p}_{i_{m+1}}) = \int_0^{\hat{p}_{i_{m+1}}} \beta_{n_{S0}+n_{Si},\, n_{F0}+n_{Fi}}(p)\, dp$.
- If i is in the bottom group, $s_i = \text{Prob}(p_i \geq \hat{p}_{i_m}) = \int_{\hat{p}_{i_m}}^{1} \beta_{n_{S0}+n_{Si},\, n_{F0}+n_{Fi}}(p)\, dp$.

The scores are obtained directly using the cumulative distribution function of the beta distribution.

5. *Misclassification Minimization with Random Perturbations.* As mentioned earlier as well, another fundamental difference between our problem and the one studied by Bradlow and Wainer (1998) is that we consider a multiperiod dynamic allocation of judges. If the number of periods (i.e., judges) is reduced to two, then the myopic benchmarks are not approximations anymore and they become optimal. However, with more judges, all the algorithms described in this section incur the risk of behaving myopically, and wrongly classifying ideas as top or bottom after very few evaluations, without further investigation. Our next algorithm uses insights from the literature on genetic algorithms (Goldberg 1989, Mitchell 1996) to limit such risk for the Misclassification Minimization algorithm. (Simulations studying the impact of similar modifications on the other benchmarks are available from the authors.) Genetic algorithms ensure diversity in the searched solutions by using random mutations: At each step the proposed solution is mutated (i.e., randomly altered) with a very small probability. Mutations are used as an "insurance policy against premature loss of important notions" (Goldberg 1989, p. 14). We introduce mutations, or in our case rather perturbations, by setting the scores assigned to each idea to: $s_i = s_i^0 + \varepsilon_i$ where s_i^0 is the score assigned by the Misclassification Minimization algorithm and ε_i is a normal Random variable with mean 0 and variance $0.01/(n_{S0} + n_{Si} + n_{F0} + n_{Fi})$. This specification for ε_i was adopted based on the magnitude of the mutation rate typically used in genetic algorithms, and on recent research suggesting that decreasing the mutation rate over the number of iterations (or generations) leads to higher performance compared to a fixed mutation rate (Fogarty 1989; Hesser and Männer 1991, 1992; Bäck and Schutz 1996). Simulations available from the authors suggest that the results are robust to small variations in this specification.

6. *Maximize the Right Tail.* Our next algorithm selects the ideas that have the highest current tail probabilities, i.e., that are the most likely to have a probability higher than a given threshold p_0. The scores are given by:

$$s_i = \Pr(p_i \geq p_0) = \int_{p_0}^{1} \beta_{n_{S0}+n_{Si},\, n_{F0}+n_{Fi}}(p)\, dp.$$

We set p_0 to the $100 \cdot (1 - m/I)$th percentile of the prior distribution, such that the expected number of ideas with a probability higher than p_0 is m. For example, with $I = 100$, $m = 10$ and with a uniform prior, $p_0 = 0.9$.

7. *Random Selection.* Our last algorithm simply performs a random selection of the ideas (each idea is equally likely to be shown to the next consumer).

3. Simulations

The goals of our simulations are twofold. First, we compare the performance of the algorithms introduced in the previous section under various assumptions on the true and prior distributions of the probabilities $(p_i)_{i\in\{1,\ldots,I\}}$. Second, we attempt to understand the source of the differences in performance by studying the behavior of each algorithm. In particular, we identify some of the drivers of high performance, as well as some of the limitations of each approach. Such understanding is crucial to allow future research to further improve performance.

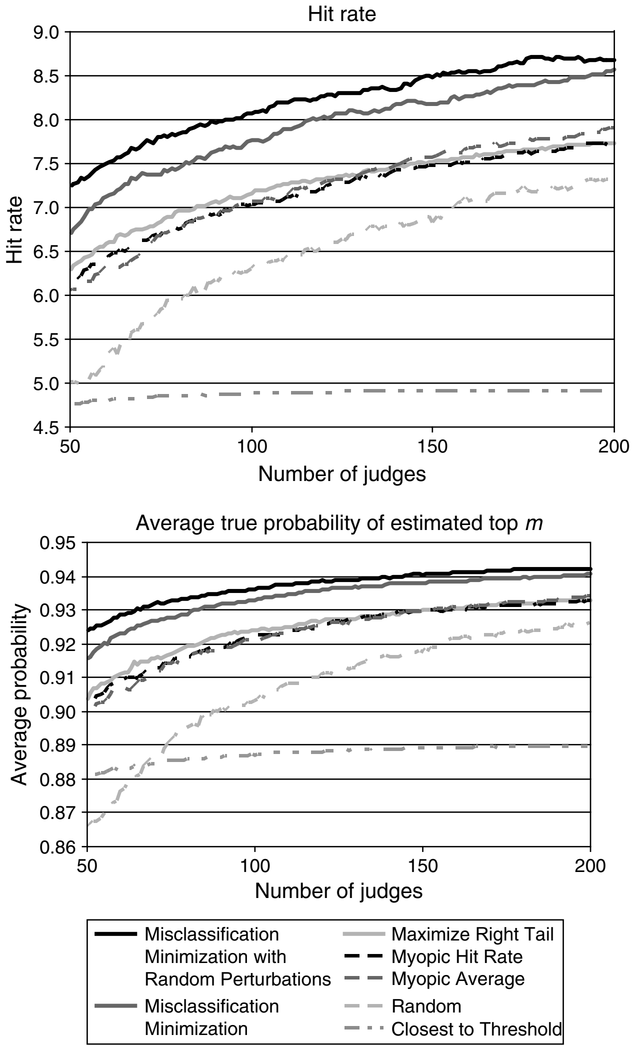
Initial Simulations

Our first set of simulations uses $m = 10$, $k = 10$, $I = 100$, $N = 200$ (i.e., the goal is to identify the top 10 ideas out of 100, and 200 consumers evaluate 10 ideas each). The true probabilities $p_1, \ldots, p_I$ are drawn from *unif* $[0, 1]$, and uniform priors are used by all algorithms ($n_{S0} = n_{F0} = 1$). Figure 1 (based on the average of 200 sets of simulations) reports the average performance as a function of the number of judges for all

[6] Bradlow and Wainer (1998) use misclassification probabilities (§§6.3 and 6.4) to determine the *number* of judges who should be allocated to each item in the rescoring phase. More precisely, they set this number proportional to the distance to the cutoff score, the coefficient of proportionality being chosen in order to minimize misclassification probability.

[7] The exact probability depends on the probability distributions associated with the entire set of I ideas. It would be challenging to estimate this exact probability adaptively without creating noticeable delays.

Figure 1 Basic Simulation Results: Misclassification Minimization with Random Perturbations Performs Best



Notes. $I = 100$, $m = 10$, $k = 10$, $N = 200$, uniform prior, $p_i \sim unif[0, 1]$.

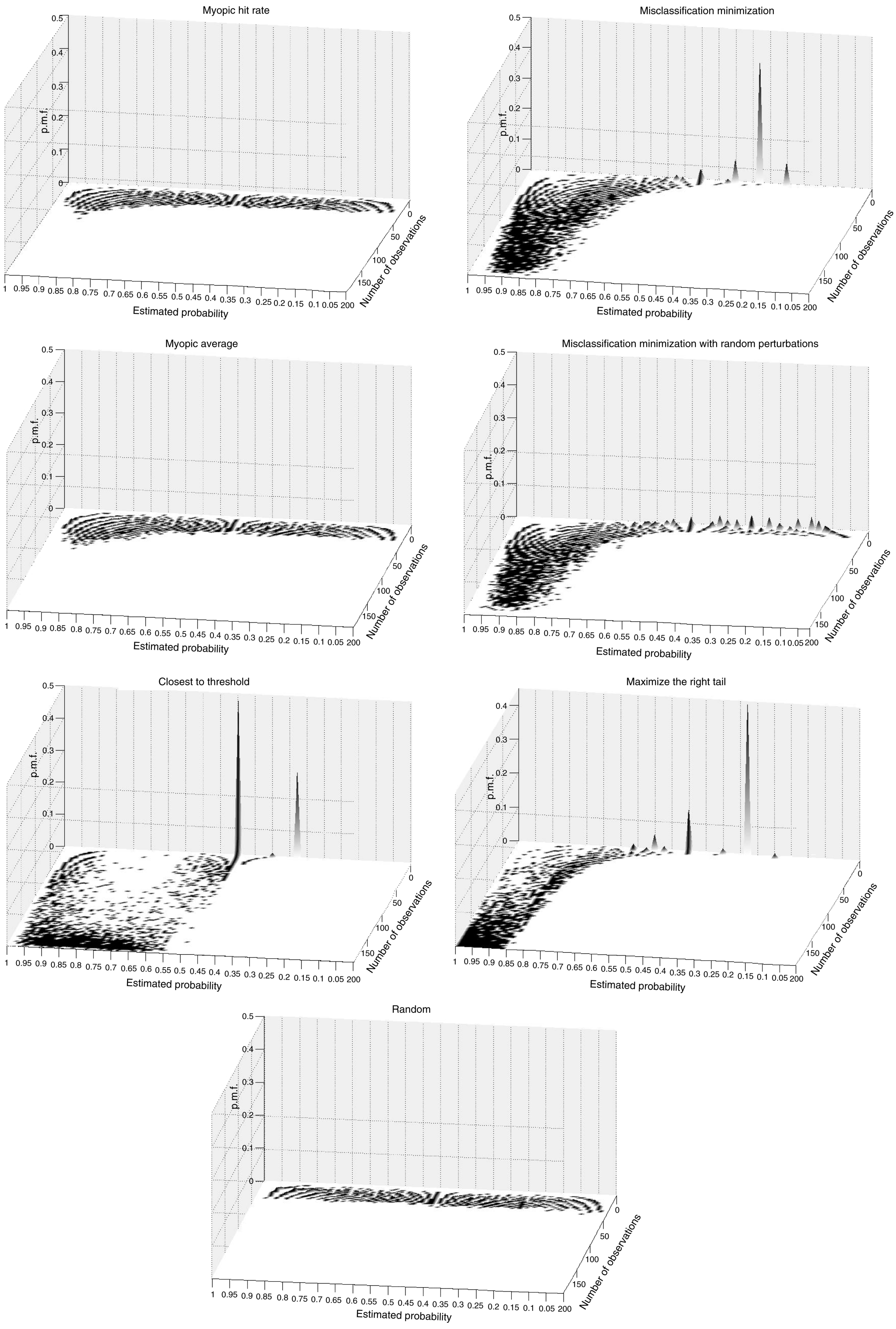
seven algorithms on both metrics (to make their reading easier, the graphs start at 50 judges). The results suggest that Misclassification Minimization with Random Perturbations performs best overall: It achieves the highest performance after 188 out of the 200 (1 to 200) possible numbers of judges on the hit rate metric, and 190 on the average probability metric. It performs significantly better (at $p < 0.05$) than Misclassification Minimization after 174 possible numbers of judges on the hit rate metric and 53 on the average probability metric. It performs significantly better (at $p < 0.05$) than all other algorithms after 183 possible numbers of judges on both metrics (all numbers of judges superior or equal to 18).[8] Note that in the limit all algorithms would converge to a perfect performance level. Hence, another way to analyze the results is to compare the speed of convergence of the different algorithms. The performance achieved on either metric by Misclassification Minimization with Random Perturbations with 167 judges (or any larger number) is higher than that obtained by Misclassification Minimization after 200 judges. The performance achieved by Misclassification Minimization with Random Perturbations on either metric with 85 judges (or any larger number) is higher than that achieved by any other algorithm after 200 judges.

We next attempt to understand the behavior of the different algorithms and the sources of the differences in performance. We consider the estimated probability associated with each idea after the evaluation by the last judge, as well as the corresponding number of observations: $\{\hat{p}_i, (n_{Si} + n_{Fi})\}_{i\in\{1,\ldots,I\}}$. We characterize the behavior of an algorithm by the frequency (across ideas and across replications) with which each pair $\{\hat{p}_i, (n_{Si} + n_{Fi})\}$ is observed. We expect high-performing algorithms to focus on ideas that are harder to classify, i.e., to lead to final states in which the estimated probabilities close to the threshold between the top and bottom groups are scoupled with the highest numbers of observations. Figure 2 plots the empirical probability mass functions of $\{\hat{p}_i, (n_{Si} + n_{Fi})\}$ for each algorithm. The x axis corresponds to $\hat{p}$, the y axis to $(n_S + n_F)$, and the z axis to the frequency with which the corresponding end state $\{\hat{p}, (n_S + n_F)\}$ is observed. We observe the following:

- The two best-performing algorithms (Misclassification Minimization and Misclassification Minimization with Random Perturbations) have the property that the number of observations is increasing with the final estimated probability up to a certain point (around 0.9, the average threshold between the top and bottom groups), after which it decreases. This suggests that these algorithms are able to focus on ideas that are harder to classify.
- Misclassification Minimization shows a peak at $\{\hat{p} = 1/3,\ n_s + n_F = 1\}$, such that 39.16% of the ideas receive one negative evaluation and are not presented to any subsequent judge. Misclassification Minimization with Random Perturbations, on the other hand, does not share this characteristic. Because the variance of the perturbations decreases with the number of observations, the perturbations mostly affect the ideas with low numbers of observations, and the two algorithms behave similarly on the other ideas. To understand the speed with which Misclassification

[8] Adding random perturbations to the other algorithms does not change the results. Misclassification Minimization with Random Perturbations performs significantly better (at $p < 0.05$) than all other algorithms with random perturbations in 189 out of 200 possible numbers of judges on the hit rate metric, and 182 on the average probability metric.

Figure 2 **Understanding the Behavior of the Algorithms by Studying the Probability Mass Functions of the Final States $\{\hat{p}, (n_S + n_F)\}$**



Myopic hit rate

Misclassification minimization

Myopic average

Misclassification minimization with random perturbations

Closest to threshold

Maximize the right tail

Random

Minimization discards ideas, note that if, for example, the threshold between the top and bottom group is 0.90 (the average threshold), the probability that an idea classified as a bottom idea and discarded after one negative evaluation has been misclassified is only: $\int_{0.9}^{1} \beta_{1,2}(p)\,dp = 0.01$. Indeed, in our simulations only 1.09% of the ideas discarded by Misclassification Minimization after one negative evaluation were misclassified (their average true probability is 0.3322). However, these errors could be easily corrected with only a few additional observations, and their accumulation has a substantial impact on performance. Indeed, the 39.16% of ideas discarded by Misclassification Minimization after one negative evaluation account for 14.81% of the misclassification errors. In contrast, the 39.16% of ideas with the lowest number of evaluations in Misclassification Minimization with Random Perturbation account for only 2.26% of the misclassification errors made by this algorithm. This illustrates the fact that random perturbations serve as an insurance policy against discarding ideas too quickly.[9]

- As predicted by the previous proposition, Myopic Hit Rate and Myopic Average behave very similarly to the random benchmark. In particular, they show only a modest increase in the number of observations for ideas with estimated probabilities around 0.90 (the average threshold between the top and bottom groups), i.e., ideas likely to change classification after only one additional evaluation.
- Closest to Threshold, because it does not take into account the variance of the beliefs on the estimated probabilities, gives rise to a bipolar distribution of the number of observations in which a large proportion (47.52%) of the ideas are not evaluated even once (peak at $\{\hat{p} = 0.5,\ n_s + n_F = 0\}$), and a small number of ideas are evaluated by almost all the judges. Once an idea is identified that has an estimated probability close to the threshold, it may be shown to all remaining judges even if other ideas have not been evaluated even once (because with a uniform prior these ideas have an estimated probability of 0.5, which is further away from the threshold).
- Maximize the Right Tail also underexplores a large proportion of the ideas (8.96% of the ideas are left unexplored and 43.94% are discarded after one negative and no positive evaluation), and focuses on ideas that are already known to have a large estimated probability. This algorithm does take variance into account—however, in a counterproductive way. In particular, an idea on which the beliefs have a high mean and low variance is actually *more* likely to be investigated further, although its classification is very likely to be correct.

Impact of the Prior and True Distributions

We now study how the true distribution of the idea probabilities and the prior Beta(n_{S0}, n_{F0}) influence the performance and behavior of the algorithms. We use a 3 (true distribution = Beta$(1,3)$, Beta$(3,1)$, or Beta$(0.1, 0.3)$) × 2 (uniform prior versus accurate prior) simulation design, in which each cell uses a setup different from the basic setup only on the true distribution of the p_is and on the parameters of the prior. A true distribution of Beta$(1,3)$ represents a context in which most ideas are of marginal value: The average probability is decreased to 1/4, and a greater proportion of ideas have low probabilities. Beta$(3,1)$ characterizes a situation in which most ideas are of high value. Beta$(0.1, 0.3)$ assumes a bimodal distribution of the ideas in which most ideas are of very low quality (60% of the ideas have a probability below 0.07) and a minority of ideas are of extremely high quality (6.5% of the ideas have a probability higher than 0.99).[10]

The simulations reveal the following:

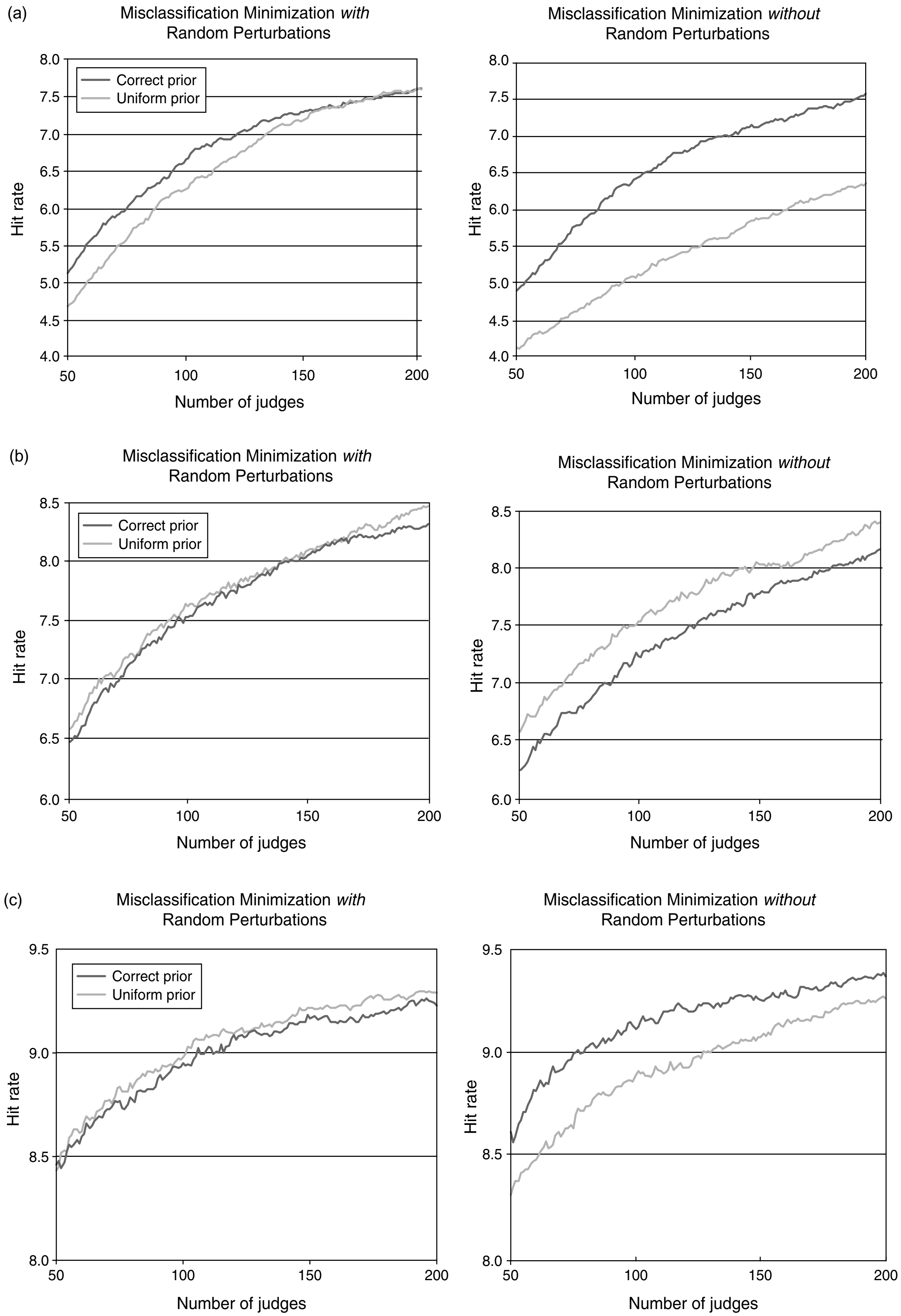
First, Misclassification Minimization with Random Perturbations continues to perform best overall. See Figures A.1–A.3 in the appendix. It performs best in 879 out of 1,200 possible comparisons (1 to 200 judges × 3 true distributions × uniform versus accurate prior) on the hit rate metric and 815 on the average probability metric. It performs significantly better (at $p < 0.05$) than Misclassification Minimization on the hit rate metric in 335 cases and in 167 cases on the average probability metric. It performs significantly better (at $p < 0.05$) than all other algorithms on the hit rate metric in 1,041 cases and in 955 cases on the average probability metric.

Second, this set of simulations further illustrates the role of random perturbations by showing that they improve robustness to variations in the prior. See Figure 3. Whereas the difference between Misclassification Minimization under a uniform versus a correct prior is significant (at $p < 0.05$) in 452 out of the 600 possible comparisons (200 numbers of possible judges × 3 true distributions) on the hit rate metric and 282 on the average probability metric, it is significant in only 105 on the hit rate metric and in 92 on the average probability metric for Misclassification Minimization with Random Perturbations. As seen previously, Misclassification Minimization tends to discard

[9] Note that another way to prevent the algorithm from discarding ideas too fast would be to force it to collect a minimum number of observations on each idea by setting the score s_i to an arbitrarily large number if idea i has been evaluated fewer than t times. However, such algorithm would perform exactly like the random benchmark on the first $t \times I/k$ judges.

[10] The symmetric distribution Beta$(0.3, 0.1)$ implies that 49.21% of the ideas have a true probability higher than 0.99, and hence is less relevant practically.

Figure 3 Random Perturbations Improve Robustness to Variations in the Prior



Notes. (a) True distribution is Beta(3, 1). (a) True distribution is Beta(1, 3). (C) True distribution is Beta(0.1, 0.3).

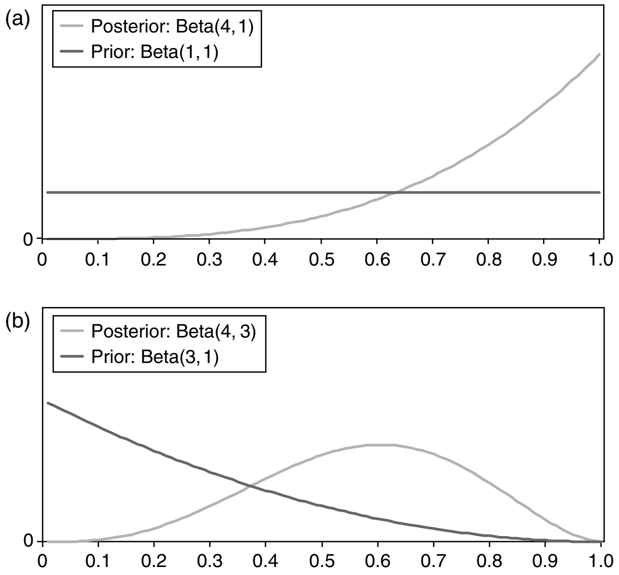
ideas after only a very few observations. The number of observations collected before discarding an idea is sensitive to the prior distribution and to the true distribution. Random perturbations decrease this sensitivity by preventing the algorithm from discarding ideas too fast.[11]

Finally, this set of simulations illustrates the fact that random perturbations serve as an insurance policy not only against classifying true top ideas as bottom ideas after too few negative evaluations, but in some cases also against classifying true bottom ideas as top ideas after too few positive evaluations. When the true distribution is Beta$(1, 3)$ and the prior is uniform, if a true bottom idea is shown to a very small set of consumers who evaluate it positively, it will have a very high estimated probability (making it likely to be classified as a top idea) and a relatively low misclassification probability (making it less likely to be further investigated). With the correct prior, such an idea would have a lower expected probability and a higher variance, and would be shown to additional consumers, who, by regression to the mean, would correct the initial evaluations. Figures 4(a) and 4(b) illustrate this effect by showing the posterior distribution of p_i after three positive evaluations, in the case of a uniform (Figure 4(a)) and a nonuniform prior Beta$(1, 3)$ (Figure 4(b)). In our simulations, under a uniform prior and a true distribution of Beta$(1, 3)$, Misclassification Minimization discards ideas with no negative evaluations after an average of 5.50 positive evaluations. This number goes up to 6.72 when random perturbations are added and to 9.32 under a correct prior. As a result, 6.17% of the final misclassification errors are made on ideas that receive only positive and no negative evaluation under a uniform prior and without random perturbations. This proportion goes down to 2.27% if random perturbations are added, and 0.54% under the correct prior.

Robustness to the Parameter k

In our simulations as well as in our field experiment, we assumed that each consumer evaluates $k = 10$ ideas and that the objective is to identify the top $m = 10$ ideas. The choice of both of these parameters raises interesting issues that may be addressed in future research. For example, increasing m decreases the likelihood of screening out high-quality ideas, but reduces the expected quality of the selected ideas. If consumers are paid a fixed fee for their participation, increasing k increases the total number of evaluations without increasing cost. On the other hand, a large value of k might lead to cognitive overload and reduce the quality of the evaluations. The number of ideas that can easily be handled by consumers is an empirical question; however, our experience suggests that consumers are comfortable with the task for values of k smaller than or equal to 12.

Figure 4 (a) Impact of Three Positive Evaluations, Uniform, Prior; (b) Impact of Three Positive Evaluations, Prior = Beta(3, 1)



We tested the robustness of our simulation results to the parameter k by replicating the initial simulations with $k = 5$ and $k = 15$, while keeping m constant at 10. The number of respondents was set, respectively, to 400 and 133 in order to maintain the total number of observations at a constant. See Figure A.4 in the appendix. The performance of all algorithms except Closest to Threshold and Maximize the Right Tail is not noticeably affected by variations in k. Although the total number of evaluations is held constant, the performance of Closest to Threshold and Maximize the Right Tail is improved when k is increased (albeit not to the level of Misclassification Minimization with Random Perturbations). Recall that both of these algorithms tend to focus exclusively on a subset of the ideas and leave the rest unexplored. Increasing k forces the algorithms to explore a wider range of ideas. For example, the proportion of ideas left unexplored by Closest to Threshold decreases from 47.52% to 38.87% when k increases from 10 to 15, and the proportion of ideas left unexplored by Maximize the Right Tail decreases from 8.96% to 0.18%.

In conclusion, our simulations suggest that Misclassification Minimization with Random Perturbations

[11] Note that when the true distribution is Beta$(1, 3)$, a correct prior actually hurts the performance of Misclassification Minimization (performance is similar under a correct versus uniform prior when random perturbations are added). This sensitivity is driven by the fact that a prior indicating that most ideas have a low probability exacerbates the algorithm's propensity to discard ideas after only very few negative evaluations.

provides the highest and most robust performance. Random perturbations serve as an insurance policy against discarding ideas too fast, and improve the robustness to changes and misspecifications of the prior. Based on these results, we will not further compare the relative performance of the algorithms. Instead, we will use Misclassification Minimization with Random Perturbations in our field experiment, and focus on convergence and convergent validity.

4. Field Experiment

We used simulations to compare and analyze the idea-screening algorithms introduced in §2 because simulations enable exact performance evaluations, replications and powerful significance tests, and the study of the impact of changes in some parameters or assumptions. However, they assumed that the true probabilities p_is were i.i.d. from a beta distribution, and that the judges' evaluations were i.i.d. from a binomial distribution. Violations of these assumptions may impact the validity of using simulations to study and compare idea-screening algorithms, as well as the validity of involving consumers in idea screening. Our field experiment attempted to address these concerns by (1) comparing the convergence observed empirically to that suggested by simulations, (2) testing the convergent validity of the proposed approach.

Design of the Experiment

This experiment consisted of two phases. Phase I allowed us to explore convergence, and Phase II to assess convergent validity. Both phases were run in collaboration with crmmetrix$^{\text{TM}}$ (www.crmmetrix.com), a marketing research company.

In Phase I, Misclassification Minimization with Random Perturbations was used to identify the top 10 out of 99 consumer-generated ideas on new cell phone features and on potential improvements to current features.[12] A random set of members of crmmetrix$^{\text{TM}}$'s online panel was invited to participate by e-mail, resulting in 195 respondents. Given the market penetration of the category, respondents were not screened. Each participant evaluated one subset of 10 ideas. In order to form a prior, we asked three pretest respondents to rate all 99 ideas on a five-point scale.[13] We fitted the probability of receiving a score of four or five with a beta distribution. Fit was maximized with Beta$(0.97, 0.76)$. We rounded these parameters and used a uniform prior of Beta$(1, 1)$.

Table 2 Convergent Validity: Top Ideas Identified in Phase I Receive Higher Scores in Phase II

Idea	Estimated group based on Phase I	Proportion of positive evaluations in Phase II (%)
"Everywhere in the world would get a good signal so you would never be out of range."	Top	77.38
"No charge for text messages and sending/receiving pics."	Top	69.41
"Have a built-in GPS navigation."	Top	66.67
"Getting Internet access would be included on the plan with no extra charge."	Top	65.88
"No need to scream into their phones anymore because the mics are much more sensitive."	Top	48.81
"It can talk and tell you who is calling and you can tell it to answer or just send them to your voice mail."	Bottom	45.88
"I would definitely improve on the Internet browser, they are too hard to use."	Bottom	37.65
"Download movies with your phone and project them on the wall so it seems like you're at the theater."	Bottom	25.88
"Ability to download software to watch DVDs from my cell phone. It would be very small with a big enough nice screen for displaying my DVDs...."	Bottom	21.18
"There would be a way to ftp data files."	Bottom	20.24

In Phase II, using a similar interface, we asked an additional 85 respondents to rate a *unique* subset of 10 ideas: 5 ideas from the top 10 identified in Phase I (4 ideas randomly selected from the top 5 and the 10th-ranked idea) and 5 ideas from the bottom group (the 11th-ranked idea and 4 bottom ideas randomly selected from those with the fewest evaluations). See Table 2 for the list of ideas tested in Phase II (the ideas were presented to the respondents in a randomized order). In a commercial application, it would be advisable to segment respondents based on demographic and usage variables, and to screen ideas at the segment level. In particular, certain ideas may be particularly popular among respondents who are the most active in the category. Hence, we do not claim that the top ideas from Phase I are the ones that would lead to the highest revenues. Recall that our goal is to study convergence and convergent validity, not to make recommendations specific to the cell phone market.

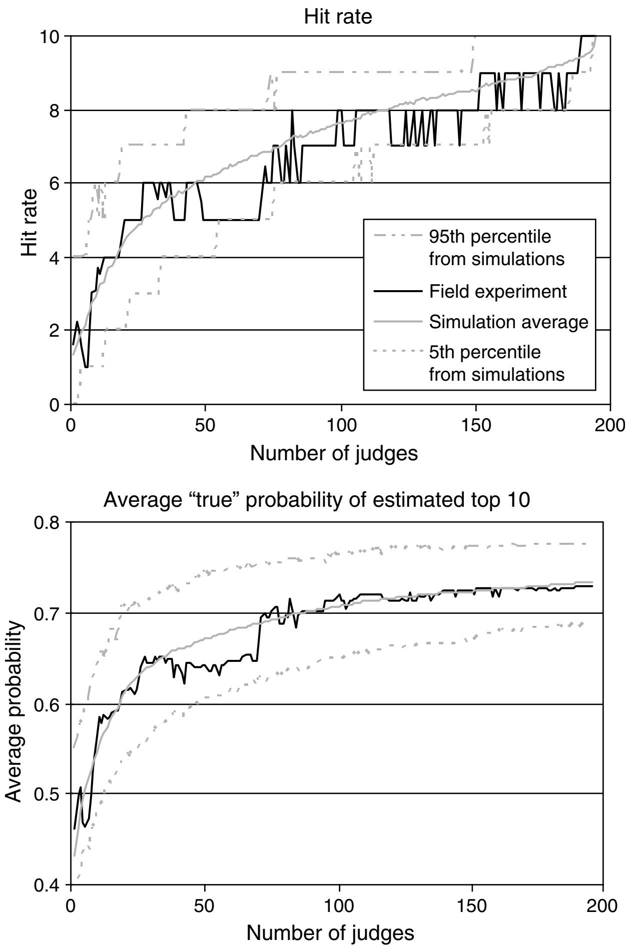
Results from Phase I: Convergence

Our goal is to assess the validity of using simulations for studying idea-screening algorithms. Because exact performance evaluation is impossible when the "truth" is unknown, we compare the pattern of *convergence* observed in the field experiment to that

[12] These ideas were prescreened to eliminate redundant and irrelevant items.

[13] The question asked of the three pretest respondents was: "Assume that you are on a new products team, trying to identify new possible features that could be included in cell phones, as well as opportunities for improving current features. Please indicate which of the following ideas you would pursue." The response scale was "definitely not," "probably not," "may be," "probably yes," or "definitely yes."

Figure 6 Convergence Pattern from Field Experiment vs. Simulations



suggested by simulations. We fitted the empirical observations from Phase I with a beta distribution (giving rise to Beta(6.2, 8.3)) and ran 200 sets of simulations using the same parameters as in the field experiment ($I = 99$, $k = 10$, $m = 10$, $N = 195$). Convergence is obtained by assuming that the probability estimates obtained after the evaluations by the last judge represent the truth, and computing performance after each intermediate number of judges. Similarly to the simulations, we define performance as hit rate (i.e., proportion of final top 10 ideas in the top 10 after judge n) or as the average true probability of the ideas estimated to be in the top 10.

We report the results in Figure 5. We see that the convergence pattern suggested by simulations is very comparable to that observed in the field. The convergence graphs from the field experiment lie within the 90% confidence bounds defined by the 5th and 95th percentiles (across the 200 replications) from the simulations.

Results from Phase II: Convergent Validity

Table 2 reports the proportion of positive evaluations obtained by each of the 10 ideas tested in Phase II. All ideas classified in the top group based on Phase I achieved higher scores than all ideas classified in the bottom group. Of all 25 possible pairwise comparisons (5 "top" ideas × 5 "bottom" ideas), 23 are significant at the $p < 0.01$ level. This suggests that adaptive idea screening using consumers based on Misclassification Minimization with Random Perturbations has good convergent validity.

5. Managerial Application: BrandDelphi^{TM}

The research reported in this paper has been applied in the brandDelphi^{TM} product (www.branddelphi.com) offered by crmmetrix^{TM} (www.crmmetrix.com). BrandDelphi^{TM} asks respondents (typically recruited from a consumer panel or a client database) to perform two tasks in sequence: (1) generate ideas on a given topic and (2) evaluate a subset of the ideas proposed by the previous participants. In the initial version of the product, ideas in the second stage were selected using a method close to random selection. The company has now adopted our Misclassification Minimization with Random Perturbations algorithm.

6. Conclusions and Opportunities for Future Research

In this paper we propose a practical tool for involving consumers in idea screening. The small set of ideas to be evaluated by each consumer is adaptively selected based on the previous evaluations. We test, analyze, and compare seven idea-screening algorithms using simulations. We study the convergence and convergent validity of the best-performing algorithm using a field experiment.

This paper constitutes only a first step towards studying algorithms for adaptive idea screening using consumers. We have already mentioned several areas for future research throughout the paper. We believe there are more. First, the performance of the different algorithms could be further compared using a field experiment. Second, interpendence between ideas could be captured by a hierarchical model in which the prior $p_i \sim Beta(n_{S0}, n_{F0})$ would be replaced with $p_i \sim Beta(\exp(\Theta_S \cdot z_i), \exp(\Theta_F \cdot z_i))$, where the z_is would be idea covariates (the priors on Θ_S and Θ_F would probably be diffuse). The covariates z_is could, for example, be dummy variables capturing the category of the idea (e.g., ideas about the size of the cell phone, about the cost of the service plan, about games and entertainment, etc.), or they could capture determinants of success identified in previous research (e.g., Goldenberg et al. 2001). Third, one could consider evaluations of ideas on multiple, potentially correlated dimensions (e.g., originality and feasibility), using, for example, a Sarmonov distribution (Danaher and Hardie 2005). Fourth, more effective algorithms may be developed using other tools, such as support vector machines or machine learning (Cui and Curry 2005, Evgeniou et al. 2005). Fifth, it would be interesting to gain a deeper understanding of the impact of endogeneity (due to adaptivity) of idea selection on the estimates of the probabilities p_is (Hauser and Toubia 2005). Sixth, the performance metrics as well as the algorithms could be generalized to give different weights to errors of different types (type I errors versus type II errors) or to errors on ideas with higher estimated probabilities. Finally, the approach of involving a large number of consumers in idea screening could be compared to that of involving a small number of experts. We hypothesize that the difference between the approaches may be understood within von Hippel's framework of "sticky information" (Von Hippel 1994, 1998; Randall et al. 2006), and that experts are more sensitive and responsive to "solution information," whereas consumers are more sensitive and responsive to "need information." (Any idea that proposes a *solution* to a *need* may be viewed as a combination of these two types of information.)

We close by noting that the framework introduced in this paper could be applied beyond idea screening. We have assumed that the judges were consumers and that the items to be judged were ideas. However, the algorithms do not rely on these two assumptions. Another possible application, relevant to the marketing academic community, could be the screening of Ph.D. applicants or job candidates. Asking all faculty members in a department to evaluate all the applications (typically around 100) may potentially result in a low response rate, and/or in noisy evaluations. The framework proposed here could be used to adaptively select a subset of the applications to be considered by each faculty member.

Acknowledgments

The authors would like to thank Oded Netzer for his comments and suggestions, as well as participants to the 2005 Marketing Science conference and the 2006 four-school colloquium at New York University. Finally, they would like to thank Hemen Patel and RK from crmmetrix for their help with our field experiment.

Appendix A. Nonbinary Evaluations of the Ideas

In this paper we consider binary evaluations by the consumers. However, finer classifications may often be useful. For example, if an idea represents an opportunity for a niche product, it is likely to be very appealing to a smaller set of consumers. It is possible to generalize the approach presented in this paper to nonbinary classifications of the ideas. For example, let us assume that we introduce three categories labeled "very good," "good," and "not good," and that our goal is to identify the 5 ideas most likely to be judged as very good and the 10 ideas most likely to be judged as good. Solving this problem could be viewed as solving two binary classification problems in parallel. A set of scores s_i^1 and s_i^2 could be assigned to each idea for each of the two problems.[14] Different criteria could then be used to select the ideas presented to the next consumer. For example, they could be the k' ideas with the highest scores with respect to the very good category and the $k-k'$ ideas with the highest scores with respect to the good category. Alternatively, they could be the k ideas with the highest average scores across the two classifications, or with the highest maximum scores.

Appendix B. Proof of the Proposition

Proposition. *Consider the set of ideas classified as top ideas that would remain classified as top after one positive or negative evaluation, and the set of ideas classified as bottom ideas that would remain classified as bottom after one positive or negative evaluation*:

$$\Omega=\left\{i\in\{i_1,\dots,i_m\},\ \frac{n_{S_0}+n_{S_i}}{n_{S_0}+n_{S_i}+n_{F_0}+n_{F_i}+1}\geq\hat{p}_{i_{m+1}}\right\}$$

$$\cup\left\{i\notin\{i_1,\dots,i_m\},\ \frac{n_{S_0}+n_{S_i}+1}{n_{S_0}+n_{S_i}+n_{F_0}+n_{F_i}+1}\leq\hat{p}_{i_m}\right\}$$

(*we assume, without loss of generality, that* $\hat{p}_{i_1}\geq\cdots\geq\hat{p}_{i_m}\geq\hat{p}_{i_{m+1}}\geq\cdots\hat{p}_{i_I}$).

Each myopic approximation assigns the same score to all ideas in the set Ω.

Proof. Let us consider an idea $i\in\Omega$. Idea i is such that the identity of the top m ideas will be unchanged after one additional evaluation on that idea. Let us first consider the myopic maximization of hit rates. We have:

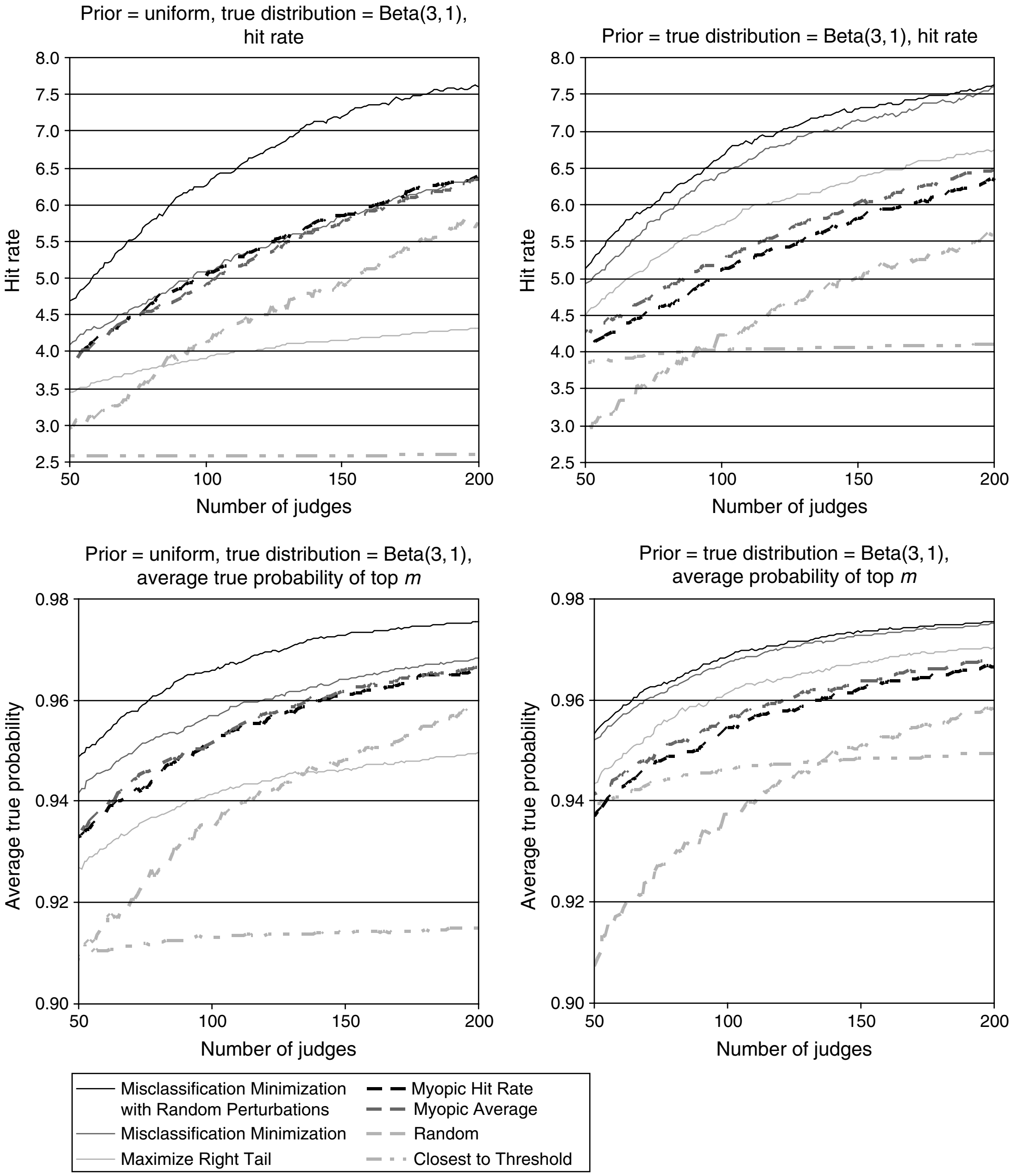
$$\begin{aligned}
s_i={}&\hat{p}_i\cdot H(n_{S1},\dots,n_{S(i-1)},n_{Si}+1,n_{S(i+1)},\dots,n_{SI},n_{F1},\dots,n_{FI})\\
&+(1-\hat{p}_i)\cdot H(n_{S1},\dots,n_{SI},n_{F1},\dots,n_{F(i-1)},\\
&\qquad n_{Fi}+1,n_{F(i+1)},\dots,n_{FI})\\
={}&\frac{n_{S0}+n_{Si}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}\cdot\frac{1}{B(n_{S0}+n_{Si}+1,n_{F0}+n_{Fi})}\\
&\cdot\int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\dots,i_m\}}\left(\prod_{j\in\{1,\dots,I\}\backslash\{i_1,\dots,i_m\}}1(p_i\geq p_j)\right)\\
&\cdot\beta_{n_{S0}+n_{S1},\,n_{S0}+n_{F1}}(p_1)\cdots(p_i)^{n_{S0}+n_{Si}}\\
&\cdot(1-p_i)^{n_{F0}+n_{Fi}-1}\cdots\beta_{n_{S0}+n_{SI},\,n_{S0}+n_{FI}}(p_I)\cdot dp_1\cdots dp_I\\
&+\frac{n_{F0}+n_{Fi}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}\cdot\frac{1}{B(n_{S0}+n_{Si},n_{F0}+n_{Fi}+1)}\\
&\cdot\int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\dots,i_m\}}\left(\prod_{j\in\{1,\dots,I\}\backslash\{i_1,\dots,i_m\}}1(p_i\geq p_j)\right)\\
&\cdot\beta_{n_{S0}+n_{S1},\,n_{S0}+n_{F1}}(p_1)\cdots(p_i)^{n_{S0}+n_{Si}-1}\\
&\cdot(1-p_i)^{n_{F0}+n_{Fi}}\cdots\beta_{n_{S0}+n_{SI},\,n_{S0}+n_{FI}}(p_I)\cdot dp_1\cdots dp_I.
\end{aligned}$$

Note that

$$\frac{n_{S0}+n_{Si}}{B(n_{S0}+n_{Si}+1,n_{F0}+n_{Fi})}=\frac{n_{F0}+n_{Fi}}{B(n_{S0}+n_{Si},n_{F0}+n_{Fi}+1)}$$

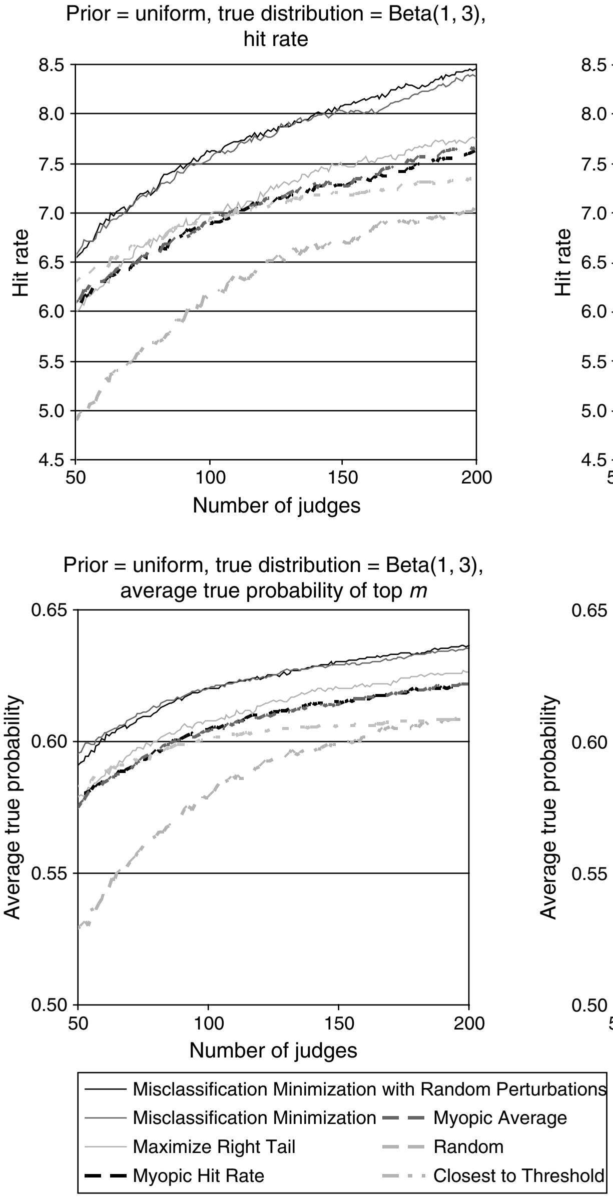
[14] Each idea would have two scores, one corresponding to the very good category and the other to the good category.

Figure A.1 Influence of the Prior Distribution



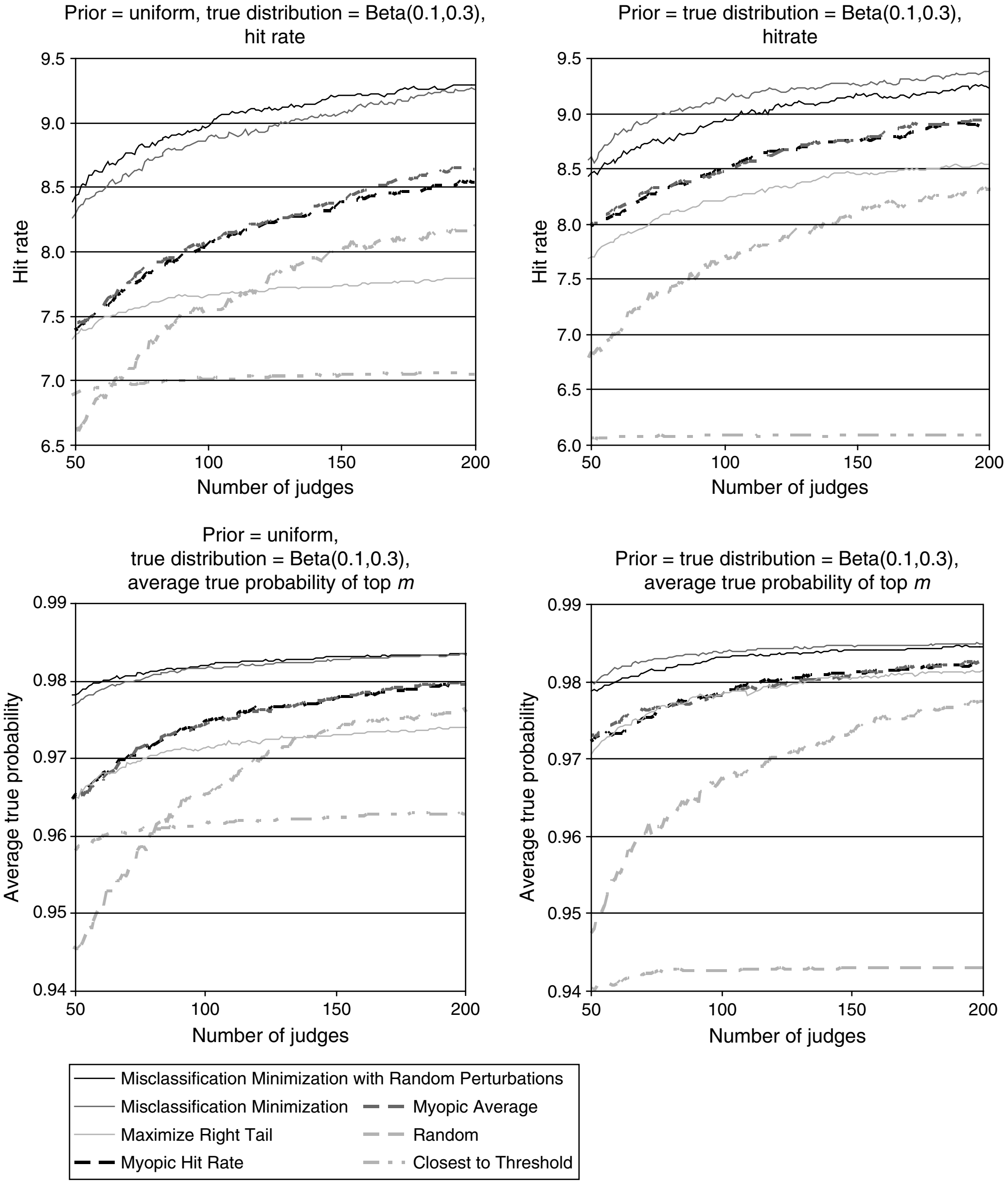
Note. True distribution is Beta(3, 1).

Figure A.2 Influence of the Prior Distribution



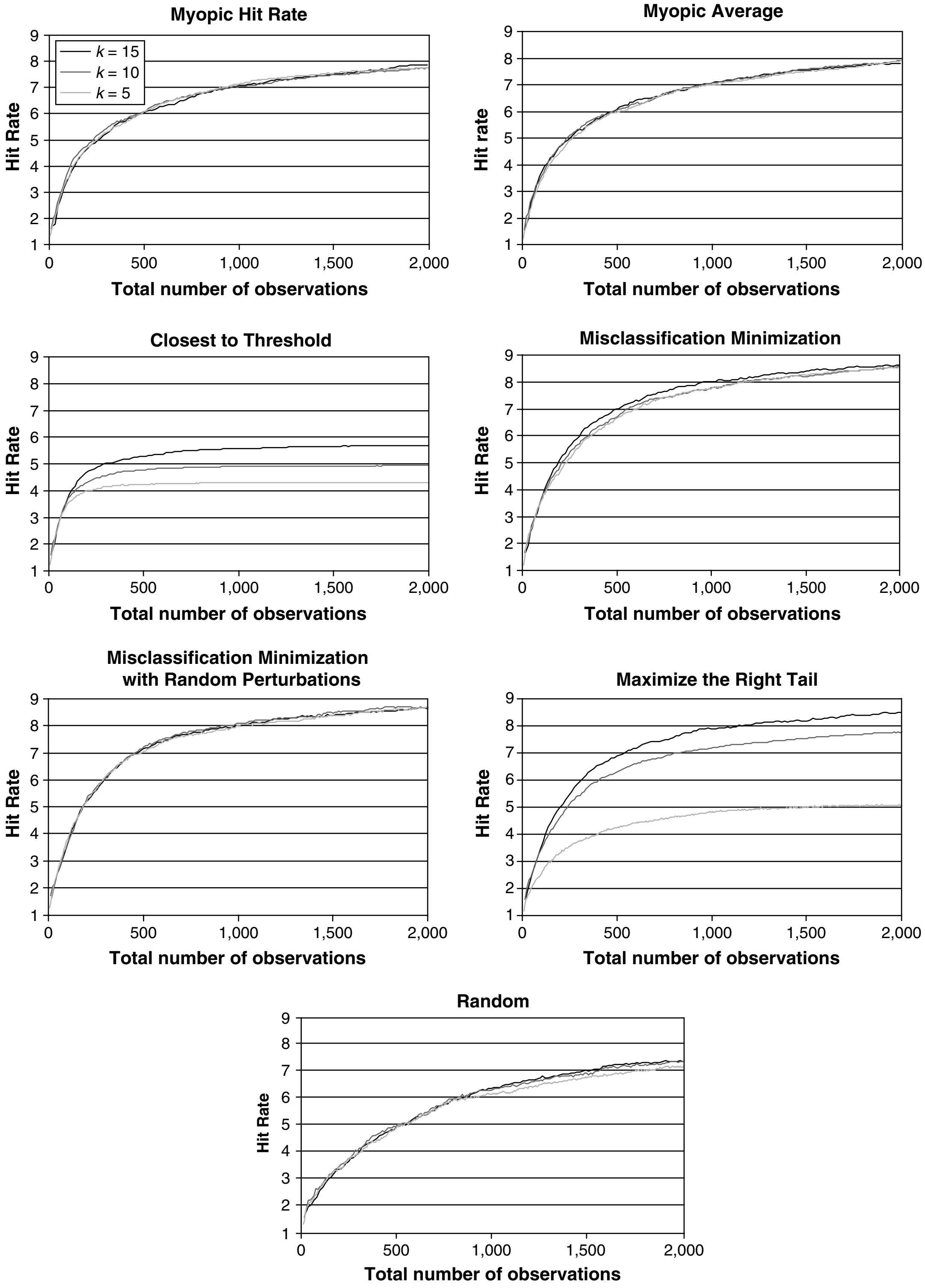
Note. True distribution is Beta(1, 3).

Figure A.3 Influence of the Prior Distribution



Note. True distribution is Beta(0.1, 0.3).

Figure A.4 Influence of the Number of Ideas Evaluated by Each Judge (k)



Myopic Hit Rate

Myopic Average

Closest to Threshold

Misclassification Minimization

Misclassification Minimization with Random Perturbations

Maximize the Right Tail

Random

because

$$
\begin{aligned}
&B(n_{S0}+n_{Si}+1, n_{S0}+n_{F0})\\
&\quad=\int_0^1 t^{n_{S0}+n_{Si}}\cdot(1-t)^{n_{F0}+n_{Fi}-1}dt\\
&\quad=\frac{n_{S0}+n_{Si}}{n_{F0}+n_{Fi}}\cdot\int_0^1 t^{n_{S0}+n_{Si}-1}\cdot(1-t)^{n_{F0}+n_{Fi}}dt\\
&\quad=\frac{n_{S0}+n_{Si}}{n_{F0}+n_{Fi}}\cdot B(n_{S0}+n_{Si}, n_{S0}+n_{F0}+1).
\end{aligned}
$$

(from integration by parts).

Hence,

$$
\begin{aligned}
s_i &= \frac{n_{S0}+n_{Si}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}\cdot\frac{1}{B(n_{S0}+n_{Si}+1, n_{F0}+n_{Fi})}\\
&\quad\cdot\int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\ldots,i_m\}}\left(\prod_{j\in\{1,\ldots,I\}\backslash\{i_1,\ldots,i_m\}}1(p_i\geq p_j)\right)\\
&\quad\cdot\beta_{n_{S0}+n_{S1},\, n_{S0}+n_{F1}}(p_1)\cdots[(p_i)^{n_{S0}+n_{Si}}\cdot(1-p_i)^{n_{F0}+n_{Fi}-1}\\
&\qquad\qquad+(p_i)^{n_{S0}+n_{Si}-1}\cdot(1-p_i)^{n_{F0}+n_{Fi}}]\\
&\quad\cdots\beta_{n_{S0}+n_{SI},\, n_{S0}+n_{FI}}(p_I)\,dp_1\cdots dp_I\\
&= \frac{n_{S0}+n_{Si}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}\cdot\frac{B(n_{S0}+n_{Si}, n_{F0}+n_{Fi})}{B(n_{S0}+n_{Si}+1, n_{F0}+n_{Fi})}\\
&\quad\cdot\int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\ldots,i_m\}}\left(\prod_{j\in\{1,\ldots,I\}\backslash\{i_1,\ldots,i_m\}}1(p_i\geq p_j)\right)\\
&\quad\cdot\beta_{n_{S0}+n_{S1},\, n_{S0}+n_{F1}}(p_1)\cdots\beta_{n_{S0}+n_{Si},\, n_{S0}+n_{Fi}}(p_i)\\
&\quad\cdots\beta_{n_{S0}+n_{SI},\, n_{S0}+n_{FI}}(p_I)\,dp_1\cdots dp_I\\
&= \int_{p_i=0}^{p_i=1}\cdots\int_{p_I=0}^{p_I=1}\sum_{i\in\{i_1,\ldots,i_m\}}\left(\prod_{j\in\{1,\ldots,I\}\backslash\{i_1,\ldots,i_m\}}1(p_i\geq p_j)\right)\\
&\quad\cdot\beta_{n_{S0}+n_{S1},\, n_{S0}+n_{F1}}(p_1)\cdots\beta_{n_{S0}+n_{Si},\, n_{S0}+n_{Fi}}(p_i)\\
&\quad\cdots\beta_{n_{S0}+n_{SI},\, n_{S0}+n_{FI}}(p_I)\,dp_1\cdots dp_I\\
&= H(n_{S1},\ldots,n_{SI},n_{F1},\ldots,n_{FI})
\end{aligned}
$$

because

$$
\begin{aligned}
&\frac{B(n_{S0}+n_{Si}+1, n_{F0}+n_{Fi})}{B(n_{S0}+n_{Si}, n_{F0}+n_{Fi})}\\
&\quad=\frac{1}{B(n_{S0}+n_{Si}, n_{F0}+n_{Fi})}\int_0^1 t\cdot t^{n_{S0}+n_{Si}-1}\cdot(1-t)^{n_{F0}+n_{Fi}-1}\,dt\\
&\quad=\int_0^1 t\cdot\beta_{n_{S0}+n_{Si},\, n_{F0}+n_{Fi}}(t)\,dt=\frac{n_{S0}+n_{Si}}{n_{S0}+n_{Si}+n_{F0}+n_{Fi}}.
\end{aligned}
$$

Hence, all ideas in Ω have the same score.

Let us now consider the myopic maximization of the average true probability of the estimated top m. The expected posterior probability associated with an idea after an evaluation is (the first and second term correspond to the updating of the estimated probability, respectively, after a positive and a negative evaluation):

$$
\begin{aligned}
&\hat{p}_i\cdot\frac{n_{Si}+n_{S0}+1}{n_{Si}+n_{S0}+n_{Fi}+n_{F0}+1}+(1-\hat{p}_i)\cdot\frac{n_{Si}+n_{S0}}{n_{Si}+n_{S0}+n_{Fi}+n_{F0}+1}\\
&\quad=\frac{n_{Si}+n_{S0}+\hat{p}_i}{n_{Si}+n_{S0}+n_{Fi}+n_{F0}+1}\\
&\quad=\frac{(n_{Si}+n_{S0})\cdot(n_{Si}+n_{S0}+n_{Fi}+n_{F0})+n_{Si}+n_{S0}}{(n_{Si}+n_{S0}+n_{Fi}+n_{F0}+1)\cdot(n_{Si}+n_{S0}+n_{Fi}+n_{F0})}\\
&\quad=\frac{n_{Si}+n_{S0}}{n_{Si}+n_{S0}+n_{Fi}+n_{F0}}=\hat{p}_i.
\end{aligned}
$$

Hence, the expected value of the objective function obtained after collecting one additional evaluation on an idea is unchanged unless the summation $1/m\sum_{i\in\{i_1,\ldots,i_m\}}^{\hat{p}_i}$ is performed over a different set $\{i_1,\ldots,i_m\}$ before and after the evaluation.

References

Bäck, T., M. Schutz. 1996. Intelligent mutation rate control in canonical genetic algorithms. *Proc. Internat. Sympos. Methodologies for Intelligent Systems*, Springer-Verlag, 158–167.

Bertsekas, D. 1995. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA.

Bradlow, E., H. Wainer. 1998. Some statistical and logical considerations when rescoring tests. *Statistica Sinica* **8** 713–728.

Cook, R. D., C. J. Nachtsheim. 1980. A comparison of algorithms for constructing exact D-optimal designs. *Technometrics* **22**(August) 315–324.

Cui, D., D. Curry. 2005. Prediction in marketing using the support vector machine. *Marketing Sci.* **24**(4) 595–615.

Dahan, E., J. R. Hauser. 2001a. Product development—Managing a dispersed process. B. Weitz, R. Wensley, eds. *Handbook of Marketing*. Sage Publications Inc., Thousand Oaks, CA.

Dahan, E., J. R. Hauser. 2001b. The virtual customer. *J. Product Innovation Management* **19**(5) 332–354.

Danaher, P. J., B. G. S. Hardie. 2005. Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. *Amer. Statistician* **59**(November) 4.

De Bono, E. 1970. *Lateral Thinking: A Textbook of Creativity*. Ward Lock Educational, London, UK.

De Bono, E. 1985. *Six Thinking Hats*. Little, Brown, Boston, MA.

Dennis, A. R., J. S. Valacich. 1993. Computer brainstorms: More heads are better than one. *J. Appl. Psych.* **78**(4) 531–537.

Evgeniou, T., C. Boussios, G. Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Sci.* **24**(3) 415–429.

Federov, V. V. 1972. Translated and edited by W. J. Studden, E. M. Klimko, eds. *Theory of Optimal Experiments*. Academic Press, New York.

Fogarty, T. 1989. Varying the probability of mutation in the genetic algorithm. M. Kaufmann, ed. *Proc. 3rd Internat. Conf. Genetic Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 104–109.

Forbes. 2005. Why companies need your ideas—How they're tapping customers to develop new products. (February 14) 78–86.

Gallupe, B. R., L. M. Bastianutti, W. H. Cooper. 1991. Unblocking Brainstorms. *J. Appl. Psych.* **76**(1) 137–142.

Gallupe, B. R., A. R. Dennis, W. H. Cooper, J. S. Valacich, L. M. Bastianutti, J. F. Nunamaker. 1992. Electronic brainstorming and group size. *Acad. Management J.* **35**(2) 350–369.

Gelman, A. B., J. S. Carlin, H. S. Stern, D. B. Rubin. 1995. *Bayesian Data Analysis*. Chapman & Hall/CRC, New York.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley, Reading, MA.

Goldenberg, J., D. Mazursky. 2002. *Creativity in Product Innovation*. Cambridge University Press, Cambridge, UK.

Goldenberg, J., D. R. Lehmann, D. Mazursky. 2001. The idea itself and the circumstances of its emergence as predictors of new product success. *Management Sci.* **47**(1) 69–84.

Goldenberg, J., D. Mazursky, S. Solomon. 1999a. Toward identifying the inventive templates of new products: A channeled ideation approach. *J. Marketing Res.* **36**(May) 200–210.

Goldenberg, J., D. Mazursky, S. Solomon. 1999b. Creativity templates: Towards identifying the fundamental schemes of quality advertisements. *Marketing Sci.* **18**(3) 333–351.

Gordon, W. J. J. 1969. *Synectics*: *The Development of Creative Capacity*. Collier-Macmillan, London, UK.

Hauser, J. R., O. Toubia. 2005. The impact of utility balance and endogeneity in conjoint analysis. *Marketing Sci.* **24**(3) 498–507.

Hauser, J. R., G. Tellis, A. Griffin. 2006. Research on innovation: A review and agenda for *Marketing Science*. *Marketing Sci.* **25**(6).

Hesser, J., R. Männer. 1991. Towards an optimal mutation probability in genetic algorithms. *Proc. 1st Parrallel Problem Solving from Nature*. Springer, London, UK, 23–32.

Hesser, J., R. Männer. 1992. Investigation of the m-heuristic for optimal mutation probabilities. *Proc. 2nd Parrallel Problem Solving from Nature*, Elsevier, Amsterdam, The Netherlands, 115–124.

Kuhfeld, W. F., R. D. Tobias, M. Garratt. 1994. Efficient experimental design with marketing research applications. *J. Marketing Res.* **s31**(November) 545–557.

Mitchell, M. 1996. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

Nunamaker, J. F., Jr., L. M. Applegate, B. R. Konsynski. 1987. Facilitating group creativity: Experience with a group decision support system. *J. Management Inform. Systems* **3**(4) 5–19.

Osborn, A. F. 1957. *Applied Imagination*, rev. ed. Scribner, New York.

Ozer, M. 2005. What do we know about new product idea selection. Working paper, Center for Innovation in Management Studies, City University of Hong Kong, Hong Kong.

Prince, G. M. 1970. *The Practice of Creativity; A Manual for Dynamic Group Problem Solving*. Harper & Row, New York.

Randall, T., C. Terwiesch, K. T. Ulrich. 2006. User design of customized products. *Marketing Sci.* **26**(2) 268–280.

The Economist. 2005. The rise of the creative consumer—The future of innovation. (March 12).

Toubia, O. 2006. Idea generation, creativity, and incentives. *Marketing Sci.* **25**(5) 411–425.

Urban, G. L., J. R. Hauser. 1993. *Design and Marketing of New Products*. Prentice Hall, Englewood Cliffs, NJ.

Valacich, J. S., A. R. Dennis, T. Connolly. 1994. Idea generation in computer-based groups: A new ending to an old story. *Organ. Behav. Human Decision Processes* **57** 448–467.

Von Hippel, E. 1994. “Sticky information” and the locus of problem solving: Implications for innovation. *Management Sci.* **40**(4) 429–439.

Von Hippel, E. 1998. Economics of product development by users: The impact of “sticky” local information. *Management Sci.* **44**(5) 629–644.