

READINGS PACKET
ON THE INFORMATION PUMP

Drazen Prelec

Sloan School of Management

Massachusetts Institute of Technology

February, 2001

Contents:

- Abstract
- Instructions for playing the Information Pump
- Theoretical Paper: D. Prelec, “A two-person scoring rule for subjective reports”

ABSTRACT

Product development requires in-depth, representative, and timely information about customer needs. The customers who provide this crucial information are typically paid for their time but not for the quality of the information itself. We develop formal incentives for collecting customer information relevant to the product design process and implement these incentives through a web-based information collection protocol, the Information Pump.

Most current methods, such as interviews, focus groups, and Voice of the Customer, presume a good-faith effort on the part of the customer. Adding quantitative incentives confers at least four advantages on each such method. First, incentives promote a uniform level of motivation and effort, which is especially important if the task is difficult or fatiguing. Second, incentives clearly communicate to the customers the nature of the information which is being solicited. Third, incentives allow customers to monitor their own performance as information providers and so improve over time. Fourth, incentives provide a way of identifying superior customers who can then be retained for subsequent studies.

There are two broad approaches that one could take to creating incentives for customer input. The first approach is to link compensation to the eventual success of the particular project. The drawback of the approach is that compensation is determined by events that are remote in time and over which the customer has slight control. The second approach is to "bootstrap" individual incentives by comparing the information provided by one customer against that provided by others at the same time. The prospects for this approach are greatly enhanced by the Internet, which makes possible the simultaneous elicitation and comparison of information from a large number of physically dispersed customers.

The goal of the Information Pump project is to develop and test a web-based discussion protocol for eliciting customer information. In the proposed application of the IP, customers will receive a rich description of a new product concept via their home terminals and then discuss the concept by sending messages to each other. The discussion is structured so that instead of simply voicing opinions, participants pose and answer each other's questions. The flow of discussion has some resemblance to a parlor game, in that there is an element of "challenge" and each participant accumulates a personal score. In essence, one gets credit for presenting statements that are non-redundant on what has previously been said and that are recognized as relevant (an "a-ha") by the others. The key feature of the protocol is that it provides incentive-compatible scoring of quantitative and qualitative information.

SUBJECT INSTRUCTIONS FOR THE INFORMATION PUMP

What is the Information Pump?

- The Information Pump is a web-based discussion game, in which players give their impressions and evaluations — typically of a picture, object, or a new idea, which is displayed in one of the windows on the screen.
- Today, the topic will be concept cars.

How does it differ from an ordinary group discussion or chat room?

- Instead of simply typing opinions, the players pose and answer each other's questions and get "points" according to how well they do.
- As a player, you score points both for: (1) the quality of your questions and (2) the quality of your responses to other players' questions.
- The scoring system is fine-tuned so that it pays to tell the truth (and think hard!).

What are the different roles in the game?

- There are 3-7 Encoders and one special player — the Dummy.
- The Encoders each see a photograph of a car prototype. Each Encoder gets a different photo, but all the photos are of the same model (in some cases it might be the interior of that model). For example, here are two pictures of the same prototype.



- The Dummy only gets a “clue” which is scrambled “jigsaw puzzle” assembled from pieces of photos of that model. (The Encoders also see this clue.) For example, something like this:



Basics of play

- Each set of photographs defines a game.
- A game has 10 rounds.

As Encoder, during each of the 10 rounds you will:

- Submit exactly one question to the group and then privately encode your personal correct answer to the question (Answer = TRUE or FALSE).
- Encode your personal correct answer (TRUE or FALSE) to each question submitted by another Encoders.
- Forecast the answers of the other Encoders to each submitted question using a 9-point rating scale.

As Dummy, during each round you will:

- Forecast the answers of the other Encoders to each submitted question using a 9-point rating scale.

SEQUENCE OF PLAY

- 1 At the start of the round, one of the Encoders (selected at random) types a statement and indicates whether that statement is TRUE or FALSE.
- 2 The other Encoders and the Dummy receive the statement but not the TRUE – FALSE value.
- 3A The other Encoders indicate whether the answer is TRUE or FALSE (for them).
- 3B The Encoders forecasts the opinion of other Encoders by typing an integer between '1' and '9'.
 - 9 means high confidence that everyone else has encoded the statement as TRUE,
 - 5 means that it's equally likely that others have encoded the statement as TRUE or FALSE.
 - 1 means high confidence that everyone else has encoded the statement as FALSE.
- 3C The Dummy forecasts the encoded answers using the same 9 point scale.
- 4 The round is scored like this:

The Dummy collects points according to the accuracy of his forecast of the encoded answers.

The Encoders collect points both for their answer and their forecast.

The answer collects points to the extent that the other Encoders' forecasts are better than the Dummy's forecast (i.e., the Encoders can figure out your answer but the Dummy cannot!).

The forecast collects points according to how accurately it predicts the answers of the other Encoders.
- 5 The TRUE-FALSE value and the scores for the round are displayed in the Log panel.

The cycle repeats unless this is the last round, in which case a new game starts.

INSTRUCTIONS AND TIPS ABOUT STRATEGY

There are three types of “moves” in this game:

- SUBMITTING A QUESTION (Encoders only)
- ENCODING AN ANSWER (Encoders only)
- FORECASTING ANSWERS (Encoders and Dummy)

Now we will describe in more detail the rules for inputting each move.

1 YOU ARE WRITING AND SUBMITTING A QUESTION

Rules

At the start of the round, you will notice a “Draft your feature here” window. You can start composing your statement here. When it is your turn to submit a feature, the contents of the draft window will be copied into a “Please enter feature...” window, which will pop up on your screen:

The screenshot shows a window titled "Please enter feature ...". Inside, there is a text input field containing "This is a car for gangsters". Below the input field, there are two radio buttons labeled "true" and "false". Underneath, it says "Others will think this statement is:" followed by a forecast rating scale. The scale has 9 columns labeled "1(false)", "2", "3", "4", "5", "6", "7", "8", and "9(true)". The rows are labeled "true" and "false". The values in the cells are as follows:

	1(false)	2	3	4	5	6	7	8	9(true)
true	-70	-40	-22	-10	0	8	15	20	25
false	25	20	15	8	0	-10	-22	-40	-70

At the bottom of the window, there is a timer that says "19 seconds left.".

You can finish editing your statement in the white area, and when you are satisfied with it you indicate TRUE or FALSE by clicking on the button below (ignore for the moment the forecast rating scale).

The horizontal bar at the bottom of the screen gives the remaining time — when the time is up, the line flashes. In that case, you will still be able to finish your statement. This is just a suggestion designed to keep the game moving along at a reasonable pace.

Objective

You should write statements whose TRUE-FALSE values might be figured out by other Encoders but not by the Dummy. You are free to write anything you like. Note, however, that both the Dummy and the Encoders will have a running log of all previous submitted statements and their TRUE-FALSE values, so that to keep the Dummy “in the dark,” you should:

- a) vary statements having TRUE and FALSE values,
- b) avoid repetitive statements, whose TRUE-FALSE values can be deduced from earlier statements.

In essence, you get points for presenting statements that are “fresh” given what others have already said, and that are recognized as relevant — as “a-ha’s” — by others.

-> Tip 1

If you feel that you have run out of statements, don't give up. BE CREATIVE! Remember — on average you cannot lose no matter how subjective your statements, because the Encoders' forecasts should be at least as good as the Dummy's (on average).

-> Tip 2

Here are some types of statements you might consider,

Style / design of the vehicle

Image of the vehicle or the company

Type of person who designed it

Type of person who would own it

Physical performance, or quality

Evaluative statements (what you like or dislike about it)

Possible uses (what it's good for, and what it's not good for)

Hypothetical scenarios in which the vehicle plays a role, e.g.,

- If it was used in a movie...
- If you had a dream about it...
- If it was a person, what kind of person it would be...
- If it was an animal, what kind of animal it would be...

2 YOU ARE ENCODING AN ANSWER TO A QUESTION SUBMITTED BY SOMEONE ELSE

Rules

When the “Please evaluate the feature...” window pops up on your screen, you should indicate whether the statement contained in the window is TRUE or FALSE by clicking on the button below (ignore for the moment the rating scale).

The screenshot shows a window titled "Please evaluate the feature ...". The statement is "This is a car for distinguished people." Below the statement are two radio buttons: "true" and "false". Underneath is the text "Others will think this statement is:" followed by a 9-point rating scale from 1 to 9. Below the scale is a payoff matrix for the user's choice (true or false) based on others' ratings (1 to 9). At the bottom, a timer shows "0 seconds left."

	1(false)	2	3	4	5	6	7	8	9(true)
true	-70	-40	-22	-10	0	8	15	20	25
false	25	20	15	8	0	-10	-22	-40	-70

Objective

The objective here is to give your true opinion, without worrying about how other people will respond. The scoring system is set up so that in the long run you will get most points by telling the truth as best you can.

3 YOU ARE FORECASTING THE ANSWERS OF OTHERS

Rules

Everyone provides a forecast for each statement by clicking on a 9-point rating scale positioned just below the words “Others will think this statement is:”

- If you are the author of the statement, then you provide your forecast in the same “Please enter feature” window (see above) where you wrote out the statement.
- If you are another Encoder, you provide your forecast in the same “Please evaluate the feature” window (see above) where you encoded your answer.

- If you are the Dummy, you provide your forecast in a window that is very similar:

Response	1 (false)	2	3	4	5	6	7	8	9 (true)
True	-70	-40	-22	-10	0	8	15	20	25
False	25	20	15	8	0	-10	-22	-40	-70

0 seconds left.

Objective

The objective of the forecast is always to accurately indicate your guess of the answers (TRUE-FALSE) of the other Encoders.

If you are an encoder you should ignore your own opinion here — it's perfectly fine to encode a TRUE answer and then forecast that others will say that the statement is FALSE.

The guess should be the probability of a TRUE value — e.g. '8' means that you feel that about 80% of the other Encoders' have given a TRUE value.

The numbers below the rating scale give the ACCURACY SCORE of for each confidence level, depending on whether the correct answer is TRUE of FALSE.

Note that:

- a) the more certain you claim to be, the bigger the stakes;
- b) as you approach the extremes ('1' and '9') the losses climb much faster than gains!

-> Tip 1

In the long run, you will score the most points if the numbers correspond to your true levels of confidence. Overusing extreme ratings is a common mistake in this game. You should reserve '1' and '9' only for those occasions when you are truly certain of the statement's TRUE-FALSE value.

-> Tip 2

If you are an encoder you don't need to follow your own opinion of the TRUE – FALSE value — it's perfectly fine to encode a TRUE answer for yourself and then forecast that others will say that the statement is FALSE.

-> Tip 3

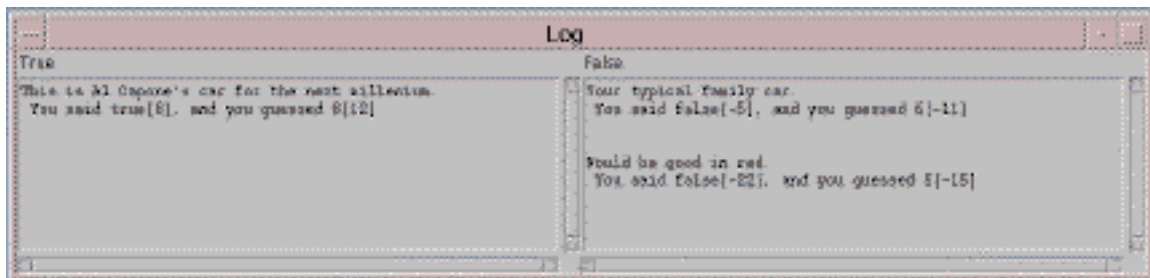
If you are the Dummy you can try to infer the TRUE-FALSE values by consulting the Log Window, which gives a record of each previous statement and how it has been answered by the Encoder who created it. A statement that is similar to an earlier one might get answered in a similar way.

4 REVIEWING THE LOG WINDOW

The log window contains two columns, a Left column for statements that were encoded as TRUE by their author, and a RIGHT column for FALSE statements.

Below each statement is a line that records how well you did when that statement was played. If you are an Encoder, then it will remind you of your answer and your forecast.

Suppose, for example, that your log window shows:



The window indicates that the statement,

“This is Al Capone’s car for the next millennium”

was encoded TRUE by the author of the statement. It also shows that your answer to the statement was TRUE and that your forecast (“guess”) of others’ answers was 8. The (8) which appears after “You said true” means that the TOTAL SCORE for the is 8 points. The (12) which appears after “guessed 8” means that the TOTAL SCORE for the forecast is 12 points.

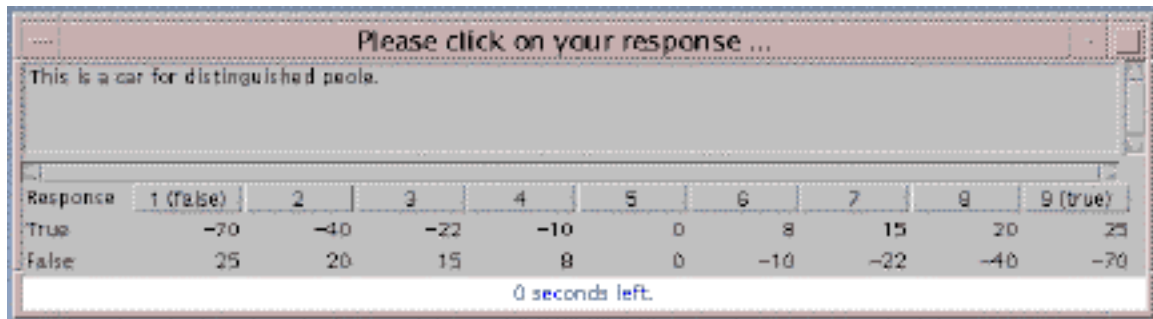
5 HOW ARE THE TOTAL SCORES COMPUTED

Dummy' forecast:

The total score for your forecast is the average accuracy score of your forecasts across all of the Encoder's TRUE-FALSE answers.

For example, if your forecast is "2" and of the three Encoders, 1 answers TRUE and 2 answer FALSE, then your total score is zero because:

$$\text{AVERAGE OF: } (-40) + (+20) + (+20) = 0.$$



Response	1 (false)	2	3	4	5	6	7	8	9 (true)
True	-70	-40	-22	-10	0	8	15	20	25
False	25	20	15	8	0	-10	-22	-40	-70

Encoder's forecast:

The total score for your forecast is computed in exactly the same way, except that at the end, the Dummy's score is subtracted (to make the game a bit more "fair"). You don't need to worry about what the Dummy will do, but you should just be aware that a positive total score means that your forecast was better than the Dummy's while a negative total score means that it was worse than the Dummy's.

Encoder's answer

The total score for your answer equals the average accuracy score of the other Encoder's with respect to your answer, minus the Dummy's accuracy score with respect to your answer.

Hence a positive total score for your answer means that other Encoder's were better able to forecast your answer than did the Dummy.

NOTE: All of these scores represent averages, hence they will not necessarily correspond to the entries on the accuracy score rows given in the window.

A two-person scoring rule for subjective reports¹

Drazen Prelec

Marketing Center

The Sloan School, MIT

Current Version: February, 2001

¹Support by the MIT Center for Innovation in Product Development, and the National Science Foundation is gratefully acknowledged.

ABSTRACT

Phenomenological data are a research staple for many disciplines. In marketing research, consumers give their impressions of product concepts, advertising materials, smells and flavors of new foods, and so forth. In linguistics or anthropology native "informants" talk about the rules and practices of their cultures. In all such situations it is standard for the scientist to regard him- or herself as a complete outsider—a "Martian"—who merely records what people say without attempting to correct them, train them, or reward them. This paper describes "bootstrap" scoring rules than can assesses the quality of such subjective judgments using the subjective judgments of other persons as the only criterion of judgmental accuracy. Scoring rules of this form can serve as a motivational incentive, as a means of identifying superior subjects, and, most importantly, as a way for subjects to teach themselves to become experts in a new and unfamiliar domain.

The main idea can be described in context of a two-person wine tasting. Irving and Jane are tasting a sample of wine, their ratings described by random variables i and j , distributed as $p(i,j)$. Irving rates this sample as i° , from which he makes a probabilistic estimate of j , $p(j | i^\circ)$. Likewise, Jane's rating is j° , and her estimate of Irving's rating is $f(i | j^\circ)$. Think of them as sitting in two separate rooms where they cannot communicate except by computer. Irving tastes the wine and writes down his judgment i , which may or may not correspond to true i° . This judgment is not shown to Jane, who instead receives from Irving a long multiple-choice question in which the "correct" answer is surrounded by many

false decoys. Jane makes two guesses about Irving's answer, once before she tastes the wine (the initial guess) and once after tasting the wine (the revised guess). The guesses are given in the form of probability distributions, and scored as log of the probability assigned to Irving's actual judgment. Irving's score, however, is the difference between Jane's revised score and her initial score. Therefore, his objective is *to maximize the improvement in Jane's ability to infer his judgment, from before to after tasting the wine*. We prove (using game-theoretic arguments) that this requires accurate reporting of his evaluation

Introduction

Many procedures in experimental psychology and marketing research make use of unvarnished introspective data. Product ratings, free associations to ad copy, similarity judgments, concept tests all rely on introspection. The subjects who happily provide this data appear to know what they are supposed to do, even without special instruction or training.

But not all responses are of equal quality (in the sense of fidelity to the underlying perceptual judgments). Some subject are confused about the instructions, some change their response criteria over time, and some are just lazy or anxious to complete the session as fast as possible. Again, however, we do not make a systematic effort to tell these subjects apart, nor to correct those whom we feel that they are making mistakes.

From this methodological limitation, it is tempting to conclude that in reality the distinction between good or bad introspection is moot, and that a purely subjective judgment, e.g., that a wine reminds you of *hollow corridors*, is not open to challenge. Indeed, the wine tasting setting nicely exposes the methodological pitfalls of uncontrolled introspection. There is a great deal of tantalizingly inconclusive discussion, sometimes ending in consensus and sometimes not. It is hard to know if disagreements are caused by semantics or genuine perceptual differences. Even in one's own case, it is hard to tell whether a wine attribute is being consistently applied over time.²

² The web has given rise to a tremendous new possibilities for collecting and disseminating subjective information about customer tastes and opinions.

By way of methodological contrast, let us compare this situation with procedures for eliciting subjective probability estimates. This is also a form of introspective judgment, and here, too, people are vulnerable to various types of mistakes. For instance, when we say that we are 90% confident in a proposition, our batting average will only be about 70%, as the research on overconfidence has shown. The subjective feeling of being 70% sure is just plain misreported as 90%. With training this particular bias largely disappears (e.g., Winkler, 1971).

Overconfidence cannot be inferred from a single judgment -- a track record is necessary. Nevertheless, it is possible to score individual assessments in such a way that accurate, well-calibrated probability assessments are maximally rewarded. The unique scoring scheme that accomplishes this treats the announced probability distribution over n events (p_1, p_2, \dots, p_n) as a bet, which is

The lack of performance-based incentives is recognized as a problem by leading academic researchers on recommendation / evaluation systems:

“Future systems will likely need to offer some incentive for the provision of recommendations by making it a prerequisite for receiving recommendations or by offering monetary compensation.” (Resnick and Varian, 1996).

“Eliciting feedback [e.g., about the quality of an interaction] encounters three related problems. The first is that people may not bother to provide feedback at all... People could be paid for providing feedback, but more refined schemes, e.g., paying on the basis of concurrence with future evaluations by others, would be required to assure that their evaluations were careful.” (Resnick, Zeckhauser, Friedman, and Ko, 2000).

“If future evaluation effort [for providing opinions or evaluations] were variable, pricing schemes could induce effort by rewarding players for evaluations that matched those of others. To do so, however, would encourage both collusion and a reluctance to state idiosyncratic opinions... We expect to return to the effort inducement problem in future work. One intriguing possibility for deterring collusion is that individuals could be rewarded for matching others, but all would be punished for a degree of agreement far beyond the statistical norm.” (Avery, Resnick and Zeckhauser, 1999).

scored as $\ln(p_{i^*})$ if event i^* turns out to be the true one.³ Faced with this scoring system, and endowed with a subjective probability distribution $\{p_1, \dots, p_n\}$, the probability assessor computes an ex-ante expected value of:

$$\sum_i p_i \ln(p_i) \quad (1)$$

which is maximized by setting the bets p_i equal to subjective probabilities π_i .

The logarithmic scoring rule provides an additional benefit, independent of the calibration issue. If the personal probabilities are built up from internal information relevant to the events plus a random error component, then the scoring rule will reward any mental work (attention, effort, discrimination) that effectively sifts relevant cues from background mental noise.⁴

This note describes a theoretical attempt to extend the methodological benefits of response scoring systems to situations where the only evidence of judgmental quality are the equally subjective judgments of another person (see Table 1 below). This defines the wine tasting problem, and it also includes a broad category of judgments concerning aesthetics, taste, social propriety,

³The logarithmic scoring rule is uniquely truth-inducing if: (i) there are at least three events, and (ii) the scoring function is of the form $f(p_{i^*})$, i.e., is constrained to depend only on the probability assigned to the true event (Savage, 1971). The expression in equation (1) is also called a directed divergence between probability distributions p and p (Kullback, 1954),

⁴If probabilities π'_i are derived from probabilities π_i by adding random error, then expected score will be reduced (assuming true reporting in both cases):

$$\sum_i \pi'_i \ln(\pi'_i) < \sum_i \pi_i \ln(\pi_i).$$

humor, archaeological provenance, and so on.

	Reporting of probabilities for objective events	Judgments without an objective accuracy definition
<i>General criterion:</i>	Forecasting accuracy	Interpersonal communication accuracy
<i>Incentive structure:</i>	Proper scoring rule	>> In this paper <<

Table 1

The remainder of this note will sketch and analyze a method for scoring the judgments of two persons who are given a common (or somehow related) set of experimental materials for evaluation. The scoring system is built up from the logarithmic rules that were just described, and overall success criterion is maximization of communicated information (in the sense of information theory).

The modelling does however introduce some new difficulties that should be briefly pointed out. First, because the scoring functions can only be based on the judgments and not on some external events, the entire procedure must be analyzed as a game where each of the two participants tries independently to maximize his or her own score (which depends on the actions of both of them). Moreover, the game in questions is partly cooperative and partly competitive (i.e., neither “zero sum” nor a pure coordination problem)...

Second, as is all games, each player is here essentially concerned with guessing what the other one will do. He or she will make inferences like: “Given that the object looks like X to me, the other person is likely to feel, notice, recognize it as Y, and label it as Z.” These beliefs may go no further than to ascribe the same perceptions to the other one, or, again, they may attempt to compensate for differences in perspective, attitude, discriminative ability, and so on. In any case, the subjective probabilities that we have to model must support an interlocking structure of mutual inference. The only way to do this, is through the empirically untested assumption of common prior over types.

Third, in games of hidden information it is customary to assume that people will tell the truth unless they have some positive incentive to lie (Rasmussen, 1989). This assumption would eliminate the problem as posed here. The providers of introspective data typically do not have any reason to lie; however, like the miscalibrated probability assessors described earlier, they need to be taught how to tell the truth. The assumption we will presently make, therefore, is that the “players” in the procedure will report only those aspects of perception for which there exists some positive incentive, however small. Players need an -incentive to reveal information.

Rules for the procedure

The procedure requires two participants, designated as Encoder and Decoder, and some material for evaluation (object, stimulus, ad copy, wine, etc.). The Encoder examines the material and writes down a report, designated as r^* . Optionally, he may also send a message to the Decoder, specifying something

about the structure of possible reports (e.g., all English statements of 100 characters or less, a list of mutually exclusive descriptions, etc.). The Decoder, seated in a separate room, writes down an initial probability distribution over possible reports, p_r . Taken literally, this is an impossible task, because of the huge number of possible reports. Let us suspend disbelief for the time being, and imagine that the Decoder can actually supply this distribution.

Subsequently, the Decoder receives the same materials and, using the new information, writes down a second revised probability distribution, q_r .

Each of these distributions is regarded as bet, and scored independently with the usual logarithmic rule,

$$V_{\text{Decoder}} = \ln(p_{r^*}) + \ln(q_{r^*}),$$

in case the actual report is r^* .

The Encoder's score is the difference between the Decoder's revised and initial scores,

$$V_{\text{Encoder}} = \ln(q_{r^*}) - \ln(p_{r^*}),$$

(where r^* is again the report actually written down). This scoring system instructs the Encoder to maximize the differential accuracy of the Decoder, before and after observing the materials. Ideally, the report should be surprising to a person who has not seen the materials, and obvious to someone who has.

Formally, this is a non-zero sum, non-cooperative game, with,

$$V_{\text{Encoder}} + V_{\text{Decoder}} = 2 \ln(q_{r^*}) - 0,$$

$$V_{\text{Encoder}} - V_{\text{Decoder}} = -2 \ln(p_{r^*}) - 0.$$

These payoff functions will support a correlated equilibrium (Aumann, 1974; 1987), meaning that the players condition their behavior on “signals” generated by some correlated randomization device even though these signals are unobservable and do not directly enter into the payoff functions.⁵

Modelling mutual knowledge

Uncertainty enters into this procedure at three points. First, there is uncertainty about what the first player, the Encoder, has actually perceived. Second, there is uncertainty about what the other player, the Decoder, conjectures about the materials before being allowed to see them. Finally, there is uncertainty about the perceptions of the Decoder after looking at the materials.

These uncertainties are represented by three integer-valued random variables, with values indexed by i,j,k, respectively. Two kinds of information are being bundled into these variables, information about certain enduring

⁵Aumann (1974) was the first to define correlated equilibria as well as to demonstrate the surprising fact that two person games can have correlated equilibria with payoff vectors outside of the convex hull of the payoffs generated by Nash (i.e., uncorrelated) equilibria. However, two person games have no interesting correlated equilibria for the polar cases of pure competition and pure cooperation: Two person zero sum games provide no incentive for coordinating actions, and in the cooperative case, i.e., when two players have identical payoffs, all the benefits of coordination can be achieved in pure strategies, without using the correlated signals.

characteristics of the other person, e.g., level of expertise, and information about actual perceptions. I will refer to the first kind as indicating a player's type and the second as their perceptions. Note that the situation here is not symmetric for the two players: for the Encoder, the i-variable contains information about both type and perceptions, while the Decoder, the j-variable indicates type and k-variable indicates perceptions.

We now invoke a standard modelling assumption, namely that the beliefs of each player about the other's characteristics and perceptions can be derived from a joint probability distributions over the relevant random variables. This means that there exists a probability distribution, (i,j,k) , which is common knowledge between both participants, and whose conditional distributions define all mutual inferences:

$(j,k | i)$ = Encoder's probability distribution over Decoder types (j) and perceptions (k)

$(i | j)$ = Decoder's initial probability distribution over Encoder type-perceptions (i)

$(i | k)$ = Decoder's revised probability distribution over Encoder type-perceptions (i)

This so-called common prior assumption (Harsanyi, 1967) suggests that although the two persons may know little about each other's perceptions per se, they are nonetheless in complete agreement about the implications of any particular perception by one person for the perceptions of the other one. The

informational content of every perception is common knowledge between them (Aumann, 1987).

We now make two further assumptions.

Assumption 1: If for all j, k , $(i | j, k) = (i' | j, k)$, then $i = i'$.

Assumption 2: For all i, j, k , $(i | j, k) = (i | k)$.

The first assumption says that the list of perceptions for the Encoder has been already reduced to equivalence classes of perceptions that induce the same beliefs about who the Decoder might be or what he or she might perceive. For the Encoder, to draw a distinction between i and i' would be tantamount to using a private randomization device (a coin toss), which would have zero impact on expected score, because neither the initial nor the revised probabilities of the Decoder will register the distinction. Taking advantage of the -incentive rule for reporting aspects, we assume that all such pairs i, i' have been pooled into their respective equivalence classes.

The second assumption says that conditional on (k) , (i) and (j) are independent. This would be true, for example, if the Decoder had perfect recall, in which case each perception k would be associated with exactly one type j . In other words, the Decoder does not lose any information by looking at the materials, perhaps a questionable assumption for wine tastings. Assumption 2 allows for some forgetting on the part of the Decoder, but maintains that the forgetting hasn't any information relevant to predicting (i) .

Derivation of a correlated, fully revealing equilibrium

The Encoder's strategy is a matrix $(r | i)$ specifying the probability that a particular perception i will lead to a particular report r . In the ensuing derivations, it will be convenient to notationally extend this matrix to the four-variate distribution, (r, i, j, k) ,

$$(r, i, j, k) \quad r(i) \quad (i, j, k) \quad (2)$$

whose marginal and conditional distributions are then denoted by dropping arguments, e.g. as in,

$$\begin{aligned} (i, j) &= \sum_{r, k} (r, i, j, k) = (i, j) \\ (k | j) &= \sum_{r, i} (r, i, k | j) = (k | j) \end{aligned} \quad (3)$$

We can think of the Encoder as choosing (r, i, j, k) subject to the constraint in equation (2).

Turning first to the Decoder's decision problem, the logarithmic scoring rule ensures that no matter what reporting strategy is believed to be in effect, the Decoder's stated inferences (initial and revised) will equal the conditional distributions $(r | j)$ and $(r | k)$, respectively:

$$\begin{aligned} p_r^o(j) &= \text{Arg Max}_{p_r} \left\{ \sum_i (r | j) \ln(p_r) \right\} = (r | j) \\ q_r^o(k) &= \text{Arg Max}_{q_r} \left\{ \sum_i (r | k) \ln(q_r) \right\} = (r | k) \end{aligned}$$

Under the assumption that the Decoder is betting optimally, we can now

compute the expected value for the Encoder of choosing r for perception i :

$$\begin{aligned} EV(r | i) &= \sum_{j, k} (j, k | i) \ln\left(\frac{q_r^\circ(k)}{p_r^\circ(j)}\right) \\ &= \sum_{j, k} (j, k | i) \ln\left(\frac{(r | k)}{(r | j)}\right) \end{aligned} \quad (4)$$

For the reporting strategy to be credible, it can only place positive probability on reports that maximize expected value:

$$^\circ(r | i) > 0 \quad \text{implies:} \quad r = \underset{r'}{\text{Arg Max}} EV(r' | i) \quad (4)$$

Note now that Assumption 2, $(i | j, k) = (i | k)$, also implies that $(r | j, k) = (r | k)$. Substituting $(r | j, k)$ into the expected value formula, and applying Bayes' rule yields,

$$\begin{aligned} EV(r | i) &= \sum_{j, k} (j, k | i) \ln\left(\frac{(r | j, k)}{(r | j)}\right) \\ &= \sum_{j, k} (j, k | i) \ln\left(\frac{(k | r, j)}{(k | j)}\right) \end{aligned} \quad (5)$$

This expression divides into two terms, of which only the first one depends on the report, r :

$$EV(r | i) = \sum_k (j, k | i) \ln(k | r, j) - \sum_k (j, k | i) \ln(k | j) \quad (6)$$

Because the contribution of the second half of Equation 6 to expected score is a constant, the Encoder is only concerned with maximizing the first half, which we introduce into equation (4) in a slightly rewritten form:

$$p(r|i) > 0 \quad \text{implies:} \quad r = \underset{r'}{\text{Arg Max}} \sum_j p(j|i) \sum_k p(k|i,j) \ln(p(k|r',j))$$

The expression,

$$\sum_k p(k|i,j) \ln(p(k|r,j))$$

looks like another instance of logarithmic scoring (and indeed is maximized for $p(k|r,j) = p(k|i,j)$). The difference here is that the Encoder doesn't directly announce $p(k|r,j)$; instead he or she needs to adopt a reporting strategy for which $p(k|r,j) = p(k|i,j)$ automatically obtains. Recalling now that every perception i is associated with a distinct distribution $p(k|i,j)$ (by Assumption 1), it follows that only a one-to-one map of perceptions into reports will support the desired equality of $p(k|r,j)$ and $p(k|i,j)$, for all j,k :

$$p(r|i) > 0 \quad \text{implies:} \quad p(r|i) = 1 \quad \text{and} \quad p(k|r,j) = p(k|i,j).$$

An implication of this is that if a distinction between i and i' is relevant to at least one type of Decoder, it will be included in the report. For instance, if the Decoder is either expert or novice, $j \in \{E,N\}$, and if:

$$\begin{aligned} p(k|i,N) &= p(k|i',N), \quad \text{for all } k \\ p(k|i,E) &\neq p(k|i',E), \quad \text{for some } k \end{aligned}$$

then the distinction between i and i' will be noted.

Expected score for the Encoder in equilibrium

To compute expected score for a fully revealing strategy, we identify (r) with (i) in equation (6) to obtain,

$$EV_{reporter}^{\circ}(i) = \sum_{j, k} (j, k | i) \ln\left(\frac{(k | i, j)}{(k | j)}\right) \quad (7)$$

The Encoder's expected score is therefore the expected information between the perceptions (i) and (k) of the two participants, conditional on Decoder type (j) . Equivalently, it is the reduction in uncertainty of (k) that (i) contributes to (j) .

Some problems

The procedure described here is problematic in several respects. As noted already, it requires elicitation of impossibly detailed probability distributions. In itself, this problem can be fixed eliciting both the report and the corresponding probabilities sequentially, in a series of smaller, more manageable steps. Such a procedure would instruct the Encoder to present the information in the form of series of statements that could be either true or false, and then elicit the probabilities for the binary statements one by one.

Psychological considerations introduce more serious problems. First, the gaming aspect of the procedure is distracting, and invites irrelevant "outguessing" strategies. Ideally, subject would understand that the game does

not allows for any systematic advantage to be gained by clever strategy, and that the best thing to do is to just attend to the evaluation materials. Better yet, the procedure could be run without the subjects aware that they are playing against a human opponent.

Second, there is little reason to be confident in the Bayesian common prior model for generating mutual inferences, in this unusual setting. People do not properly update probability distributions even in simple tasks -- why should they do that here? On the other hand, this issue is only relevant to the performance of the Decoder. It is possible that the procedure is robust with respect to fairly gross departures from normative probability assessments.

Concluding remarks

One can view this game as providing an alternative to established group judgment procedures, such as focus groups and "Delphi". Unlike focus groups, the method provides feedback that indicates when statements are making sense (i.e., when they are not random). Unlike the Delphi method, where a group consensus is essentially forced through iteration, the feedback here does not invite participants to bias their judgments toward some group mean, nor does it even require a consensus about a common group vocabulary. Finally, it would be interesting to see if the method can be used to bootstrap ad-hoc expert teams for new domains (which is often the problem in marketing). Normally, we conceive of expertise or "connoisseurship" as something you are taught by experts, which then begs the question of where the original experts got *their* expertise. With this scoring system we can take people who possess some *latent*

expertise, of which they are not fully aware, and let them train themselves, through trial and error, to become actual experts, i.e., able to recognize and communicate what they perceive and feel.

References

- Aumann, R. J. (1974). "Subjectivity and correlation in randomized strategies," Journal of Mathematical Economics, 1, 67-69.
- Aumann, R. J. (1987). "Correlated equilibrium as an expression of Bayesian rationality," Econometrica, 55, 1-18.
- Avery, Chris, Resnick, Paul, and Zeckhauser, Richard. "The Market for Evaluations." American Economic Review. 89(3): 564-584.
- Harsanyi, J. C. (1967). "Games with incomplete information played by 'Bayesian players,'" Management Science, 14, 159-189, 320-334, 486-502.
- Kullback, S. (1954). Information Theory and Statistics, New York: Wiley.
- Prelec, D. (1988). Introspection and Communication: A Game-theoretic Approach. Harvard Business School Working Paper #88-032.
- Rasmussen, E. (1989). Games and Information. Cambridge, Mass.: Basil Blackwell.
- Resnick, Paul, and Varian, Hal. "Recommender Systems," introduction to special section of Communications of the ACM, 1996, vol (39), pp. 87-93.
- Resnick, Paul, Zeckhauser, Richard, Friedman, Eric, and Kuwabara, Ko. "Reputation Systems." To appear in Communications of the ACM, December 2000.
- Savage, L. J. (1971). "Elicitation of personal probabilities and expectations," Journal of the American Statistical Association, 66, 783-801.
- Winkler, R. L. (1971). "Probabilistic prediction: Some experimental results," Journal of the American Statistical Association, 66, 675-685.