

DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

# CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

Are People Bayesian?  
Uncovering Behavioral Strategies

Mahmoud A. El-Gamal

David M. Grether

Forthcoming: *Journal of the American Statistical Association*



**SOCIAL SCIENCE WORKING PAPER 919**

February 1995

# Are People Bayesian? Uncovering Behavioral Strategies

Mahmoud A. El-Gamal and David M. Grether

## Abstract

Economists and psychologists have recently been developing new theories of decision making under uncertainty that can accommodate the observed violations of standard statistical decision theoretic axioms by experimental subjects. We propose a procedure which finds a collection of decision rules that best explain the behavior of experimental subjects. The procedure is a combination of maximum likelihood estimation of the rules together with an implicit classification of subjects to the various rules, and a penalty for having too many rules. We apply our procedure to data on probabilistic updating by subjects in four different universities. We get remarkably robust results which show that the most important rules used by the subjects (in order of importance) are Bayes's rule, a representativeness rule (ignoring the prior), and to a lesser extent, conservatism (over-weighting the prior).

*Keywords:* Classification, mixture models, probability assessments, learning.

*JEL classification numbers:* 026,211,215

# Are People Bayesian?

## Uncovering Behavioral Strategies

Mahmoud A. El-Gamal and David M. Grether \*

### 1 Introduction

The economic theory of decision making under uncertainty has been seriously challenged by a series of discoveries of violations of that theory by experimental subjects. The paradoxes of Allais (1953) and Ellsberg (1961) are among the earliest examples, but recently, psychologists have added several others. Some violations of statistical decision theory are studied in Kahneman and Tversky (1979), Lichtenstein and Slovic (1971), Grether and Plott (1979), to name but a few. Part of the economists' response to these developments has been the introduction of a number of new theoretical models of decision making designed to be consistent with some of the reported violations of expected utility theory. Some examples are Machina (1982), Loomes and Sugden (1987), Bell (1982), Chew (1983), (Quigen (1982), Yaari (1987), and Kahneman and Tversky (1979).

While economists have been introducing new models of individual decision making, psychologists have developed a number of heuristic explanations of specific individual behaviors. For some of those models see Lichtenstein and Slovic (1971), Goldstein and Einhorn (1987), Bostic et al. (1990), Loomes et al. (1989), Tversky et al. (1988), Mellers et al. (1992), and Birnbaum et al. (1992). Research on judgement of probabilities has produced an array of heuristics which individuals use in different circumstances (c.f. Tversky and Kahneman (1974)). Prominent heuristics were proposed in Kahneman and Tversky (1972), Tversky and Kahneman (1972), Edwards (1982), and Lichtenstein and Slovic (1971). Recent research (e.g. Payne (1982), Gigerenzer et al. (1991)) suggests that the dependence of such heuristics on the specific context of the decision making experiment is not fully understood. We alert the reader that in our empirical results,

---

\*We acknowledge financial support from NSF grant #SBR-9320497 to the California Institute of Technology. We thank the previous Editor (R. J. A. Little) and an anonymous Associate Editor of JASA, as well as an anonymous referee for valuable comments and suggestions. We also thank participants at the ESA meetings and the Classification Society of North America meetings, and at the Econometric workshops at Arizona, Caltech, Minnesota, Northwestern, Wisconsin, Rochester, SMU, and Texas A&M for many useful comments. Any remaining errors are, of course, our own.

two of those heuristics (representativeness, and to a lesser extent conservatism), will be discovered in the data. We shall discuss those two heuristic decision rules as well as Bayes's rule in section 3. We did not find evidence for individuals using any of the other cited heuristic decision rules, and hence we shall not discuss them any further.

The literature in its current state does not support the conclusion that subjects are sufficiently homogeneous to be described by a single theory. Different subjects may use different decision rules, and if the rules they use do not yield satisfactory outcomes, they may abandon them and use different ones (c.f. Mellers et al. (1992)). In this paper, we devise and use a general estimation/classification procedure which uncovers the most likely collection of rules that experimental subjects use.

The remainder of the paper will proceed as follows. In section 2, we describe the collection of experiments that we analyze and present an overview of the data. In section 3, we introduce a class of decision rules that reduces the computational burden to feasible levels. In section 4, we present our likelihood-based estimation/classification procedure for the particular application at hand, and motivate a particular penalty function for the number of classes allowed. In section 5, we discuss the methods used for implementing our procedure in our application, and discuss the results that we obtain from the experiments described in section 2. In that section, we also address some of the suboptimal properties that our procedure may possess in finite samples, and discuss how they can be ameliorated by considerations of optimal experimental design. In section 6, we shall compare our results to those we may achieve using the EM algorithm. Section 7 concludes the paper.

## 2 The Experiments

The experimental data that we use in this paper were collected at four different educational institutions. Subjects were recruited from economics classes at UCLA, Occidental College, California State University at Los Angeles, and Pasadena City College. The subjects were told that they were to participate in an economics experiment, and that they would be paid for their participation. Upon arrival, the subjects were randomly divided into two groups. The procedures were identical for both groups except for the method of payment. The two groups performed the experimental tasks independently in two different rooms. Since the analysis does not require that the number of subjects, or the number of tasks per subject, remain constant across experimental sessions, those numbers were determined by the number of signed-up subjects who reported at the scheduled time, and by the time-length of the sessions.

In each room, there were three bingo cages, labeled "prior cage", cage A, and cage B. For each task presented to the subjects, a draw from the prior cage was made to determine whether cage A or cage B would be used for that task. The prior cage contained six balls numbered one through six; if one through  $m$  were drawn (where  $m$  is chosen as part of the experimental design), then we used cage A for this task, otherwise we used cage B. Therefore, the choice of  $m$  induces a prior for the subjects on whether the draws

they will observe in the current task will be from cage A or cage B. Cage A contained six balls, four labelled N and two labelled G. Cage B also contained six balls, three with each letter.

The experiment proceeded as follows. Cages A and B were placed behind an opaque screen, and a value of  $m$  was announced. The prior rules used were two, three, or four chances out of six for cage A; i.e.  $m=2, 3, \text{ or } 4$ . The prior cage was spun and a ball selected thus determining whether cage A or cage B would be used. The result of this draw was not revealed to the subjects. The selected cage was placed in the front of the room and six draws (with replacement) were performed, and the results were announced and written on a blackboard. Subjects also recorded the outcomes on their answer sheets. Subjects were then asked to indicate which of cages A and B they believed was used in generating the observations. After all subjects had indicated which cage they felt was the more likely, a new value of  $m$  was announced, and the procedures were repeated.

At the beginning of each experiment, the instructions were read and the subjects elected one person to serve as a monitor. The monitors inspected all equipment, observed the draws from the cages and, generally, checked to be sure that the experimenters were being truthful. The monitors did not communicate with the subjects outside of their duties as monitor. The monitors were guaranteed a payment at least equal to the average received by subjects in their rooms.

In one room, all subjects were payed a flat fee. In the other room, one task was selected (randomly, using a bingo cage), and subjects earned a \$10 bonus if their response was correct. A response was considered correct if the cage the subject stated was the more likely was in fact the cage from which the balls were drawn. In both treatments, subjects were not given any feedback on the correctness of their responses until the very end of the experiment, when their payoffs were computed. The sessions lasted approximately one and one half hours, and the number of decisions made by each subject ranged from 14 to 21.

The aggregated numbers of A's and B's chosen for each of the three priors and each of the 7 possible outcomes are shown in Table 4. The information in Table 4 is summarized in Figure 1, where we show the proportion of A choices for each prior and each outcome, aggregated over all eight experimental sessions. The monotonicity of the proportion of A's in both the prior in favor of cage A, and the number of N's observed, suggests that subjects are to some extent using the priors and the evidence to formulate their beliefs about the parent distribution. This is indicative of some sort of noisy Bayesian behavior, and also allows for some other decision rules. In the following section, we shall introduce a large class of decision rules (including Bayes's rule) and postulate that each subject uses a single decision rule from that class throughout the experimental session, but that they are prone to make errors and make random decisions.

### 3 A Natural Class of Decision Rules

Ignoring the order of the draws, there are seven possible outcomes (zero through six N's) and three priors, resulting in 21 possible decision situations. In each of these situations, the subjects could choose either cage A or cage B. Therefore, there are in principle  $2^{21} = 2,097,152$  possible decision rules. Since cage A has a higher proportion of N's than cage B, the outcome most strongly favoring cage B would be no N's (six G's), and the one most strongly favoring cage A would be six N's (no G's). A natural rule would be to have a cutoff number for each of the priors, such that if the number of N's exceeds that cutoff number, the rule selects cage A, otherwise it selects cage B.

To use the cutoff class of rules, we need to decide how to treat the behavior of a subject who is observed choosing cage B when some number of N's has been observed, and then choosing A when a smaller number has been observed. We shall introduce the possibility of subjects making errors (i.e. deviating from the rule). This will allow each of our decision rules to give a positive probability (likelihood) to all possible patterns of behavior. We shall assume that each subject uses a decision rule  $(c_1, c_2, c_3)$ , of the form: under prior  $i$ , choose cage A if the number of N's observed is greater than  $c_i$ , and choose cage B otherwise. With probability  $\epsilon$ , however, the subject trembles and makes a random choice. In other words, for each decision (given a prior, and a number of N's drawn) with probability  $(1 - \epsilon)$  the subject follows the rule  $(c_1, c_2, c_3)$ , and with probability  $\epsilon$  the subject chooses cage A with probability  $1/2$  and cage B with probability  $1/2$ . Notice that subjects' choices agree with the rule with probability  $1 - \epsilon/2$  and deviates from it with probability  $\epsilon/2$ . Since  $\epsilon$  is the probability of acting randomly, we call it the error rate. Now, the number of possible rules  $\{(c_1, c_2, c_3); -1 \leq c_i \leq 6; i = 1, 2, 3\}$  is  $8^3 = 512$  (where  $c_i$ 's are integers, and we use  $-1$  as the lower bound corresponding to always choosing cage A, even if zero N's were observed). With perfect foresight, we now identify the three rules in this class that will appear prominently in our empirical results.

A subject who correctly uses Bayes's rule chooses cage A if the prior in favor of cage A was  $1/3$ ,  $1/2$ , or  $2/3$ , and the number of N's was greater than 4, 3, or 2, respectively. In our notation, that means that the cutoff rule  $(c_1, c_2, c_3) = (4, 3, 2)$  corresponds to Bayes's rule. A second decision rule that will appear in our estimates is representativeness heuristic (Kahneman and Tversky (1972)) which identifies samples with parent distributions that coincide with them (e.g. a sample of 3 N's and 3 G's coincides with the true composition of cage B). Therefore, a subject who uses the representativeness heuristic would judge that samples of 3 N's came from cage B and samples of 4 N's came from cage A, regardless of the prior used. The representativeness heuristic combined with a cutoff rule implies a judgement in favor of A with 4 or more N's and in favor of B with 3 or less N's, resulting in the rule (333). A third class of subjects that we wish to identify are conservative Bayesians. Those subjects give more weight to the prior odds than Bayes's formula dictates. For instance, subjects using the cutoff rule (531) must observe six N's in order to pick cage A when the prior favoring A is  $1/3$ . Note that due to the discreteness of our observations, subjects could be conservative and yet use the rule (432). However, subjects using (531) are definitely conservative (see Edwards (1982)).

There are two ways in which we could introduce “learning” on the part of the subjects. One way is to allow  $\epsilon$  to decrease over time, and another is to allow the rules (cutoff numbers for different priors) used by the subjects to change over time. We do not allow either kind of learning in our estimated class of models. This restriction should not seem too strong in light of our design where the subjects were never given any feedback about the performance of their decision rule until the end of the session.

## 4 A Likelihood-Based Estimation/Classification Procedure

As stated in the previous section, we have restricted attention to a class of decision rules which can be written as  $(c_1, c_2, c_3)$  where  $c_i$  is the cutoff rule used when prior  $i$  is induced. We assume that each of our subjects uses one such rule  $(c_1^s, c_2^s, c_3^s)$  from the class  $\mathfrak{C} = \{(c_1, c_2, c_3) : -1 \leq c_i \leq 6; i = 1, 2, 3\}$ . Different subjects may be using different rules. We further assume that the error rate  $\epsilon$  is the same for all subjects, and all tasks. Each decision rule  $c \in \mathfrak{C}$  and error rate  $\epsilon$  define a probability function  $f^{c,\epsilon}: X \rightarrow [0, 1]$ , where  $x \in X$  is a collection of triples (prior, number of N’s observed, and choice(A or B)). For a subject  $s$ , given a sequence of observations  $x_1^s, \dots, x_{t_s}^s$ , where  $x_\tau^s = (p_\tau, N_\tau, a_\tau)$ ,  $p_\tau \in \{1, 2, 3\}$  is the prior,  $N_\tau$  is the number of N’s observed, and  $a_\tau$  is the choice (A or B) of the subject, define the variable

$$x_{c,\tau}^s = \begin{cases} 1 & \text{if } (a_\tau=A \text{ and } N_\tau > c_{p_\tau}) \text{ or } (a_\tau=B \text{ and } N_\tau \leq c_{p_\tau}); \\ 0 & \text{otherwise.} \end{cases}$$

Thus,  $x_{c,\tau}^s$  equals 1 if the subject  $s$ ’s decision on trial  $\tau$  agrees with rule  $c$ . Now define the sufficient statistic  $X_c^s = \sum_{\tau=1}^{t_s} x_{c,\tau}^s$  (the number of decisions that agree with rule  $c$ ). Then under rule  $c = (c_1, c_2, c_3)$ , and error rate  $\epsilon$ , the subject follows rule  $c$  with probability  $(1 - \epsilon/2)$ . The likelihood of  $(c, \epsilon)$  given this subject’s actions  $(x_1^s, \dots, x_{t_s}^s)$  can, therefore, easily be computed as follows:

$$f^{c,\epsilon}(x_1^s, \dots, x_{t_s}^s) = \left(1 - \frac{\epsilon}{2}\right)^{X_c^s} \times \left(\frac{\epsilon}{2}\right)^{t_s - X_c^s}.$$

Now, we observe data on  $n$  experimental subjects, with each subject  $s$  being observed over  $t_s$  tasks. If we assume that all agents are using the same rule  $c \in \mathfrak{C}$ , then we can estimate  $(c, \epsilon)$  by the maximum likelihood estimates:

$$(\hat{c}, \hat{\epsilon}) = \operatorname{argmax}_{c,\epsilon} \prod_{s=1}^n f^{c,\epsilon}(x_1^s, \dots, x_{t_s}^s).$$

If instead we assume that different agents may be using different rules, and that there are exactly  $k$  such rules  $c^1 = (c_1^1, c_2^1, c_3^1), \dots, c^k = (c_1^k, c_2^k, c_3^k)$ , let  $\delta_{ij}$  be 1 if agent  $i$  is using

the  $j^{\text{th}}$  rule, and 0 otherwise. Then we can estimate  $(c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}_{1 \leq i \leq n; 1 \leq j \leq k})$  by

$$(\hat{c}^1, \dots, \hat{c}^k, \hat{\epsilon}, \{\hat{\delta}_{ij}\}_{1 \leq i \leq n; 1 \leq j \leq k}) = \operatorname{argmax}_{c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}_{1 \leq i \leq n; 1 \leq j \leq k}} \prod_{s=1}^n \prod_{h=1}^k \left( f^{c^h, \epsilon}(x_1^s, \dots, x_{t_s}^s) \right)^{\delta_{sh}}, \quad (4.1)$$

where  $\delta_{ij} \in \{0, 1\}$ , and  $\sum_{j=1}^k \delta_{ij} = 1$ . In other words, for a given  $k$ , we find estimates  $(\hat{c}^1, \dots, \hat{c}^k, \hat{\epsilon})$  as follows:

*Algorithm A:*

- For each  $(c^1, \dots, c^k, \epsilon)$ :
  - For each individual  $s$ :
    - \* Calculate  $f^{c^h, \epsilon}(x_1^s, \dots, x_{t_s}^s)$ , for  $h \in \{1, \dots, k\}$ .
    - \* Choose  $h \in \{1, \dots, k\}$  which yields the highest  $f^{c^h, \epsilon}(x_1^s, \dots, x_{t_s}^s)$ . Call the maximal value  $f_s(c^1, \dots, c^k, \epsilon)$ . This corresponds to maximizing over the  $\delta_{sj}$ 's for individual  $s$ .
  - multiply the obtained likelihoods  $f_s(c^1, \dots, c^k, \epsilon)$  over individuals  $s \in \{1, \dots, n\}$ . Call the outcome  $F(c^1, \dots, c^k, \epsilon)$ .
- Choose  $(\hat{c}^1, \dots, \hat{c}^k, \hat{\epsilon})$  to maximize  $F(\cdot)$  (equation (4.1)).

In general, the last step of the above algorithm would be implemented by calling a general purpose multivariate maximization routine. In our application, there are finitely many  $k$ -tuples of rules to check, and for each  $k$ -tuple, each individual's contribution to the likelihood function can be maximized by assigning them to the rule which minimizes the number of deviations of that individual's actions from the assigned rule. The estimate of  $\epsilon$  is then easily calculated as twice the proportion of overall deviations.

Notice that by following *Algorithm A*, our problem (for a fixed  $k$ ) becomes a simple likelihood maximization one, with the number of parameters ( $3k$  for the  $k$ -tuple rules +  $nk$  zeros and ones for the classifications of the  $n$  subjects) growing linearly in the number of rules, and linearly in the number of subjects. This improvement over the brute-force algorithm (searching over all  $k^n$  possible allocations of  $n$  subjects into  $k$  groups) makes our procedure rather easy to implement by invoking any of the standard multidimensional optimization subroutines generally available in mathematical and statistical packages.

In Section 6, we shall compare our approach to this model of subject heterogeneity with the more common treatment of the  $\{\delta_{ij}\}$ 's as missing data and use of the EM algorithm to obtain estimates of the other parameters (Redner and Walker (1984), Little and Rubin (1987)). The EM algorithm integrates out the  $\{\delta_{ij}\}$ 's to obtain an expected log likelihood function which is then maximized, whereas we treat them as parameters of the model over which we maximize the likelihood function. It is common to interpret the maximization of the likelihood function over those parameters as an approximation to maximizing the integral. The accuracy of this approximation improves as the number

of tasks per subject gets larger. This is in the same spirit that a maximum likelihood estimation is an approximation of the integral required to calculate a Bayes posterior. It turns out that the integral is quite simple to calculate in this class of mixture models, and we shall discuss it in more detail in Section 6.

For each  $k$ , we have shown how to estimate the  $k$  most likely rules, and the error rate  $\epsilon$ . Given  $n$  subjects, we still have to decide on a method to estimate  $k$ . Clearly, as  $k$  increases, the overall likelihood will increase, until  $k = n$ , the number of subjects. We would like to introduce a penalty for allowing too many decision rules, i.e. a penalty for  $k$  getting too large. There is a substantial literature on the problem of choosing an optimal penalty for the complexity of a model. The best known, and one of the earliest, is Akaike (1974)'s criterion (which picks the model that maximizes the maximal log likelihood less the number of parameters). Another very popular information criterion was introduced by Schwarz (1978), which picks the model that maximizes log of the maximal likelihood less the number of parameters multiplied by log of the sample size, divided by two. Many other criteria are implicit in the coding literature such as Wallace and Boulton (1968)'s Minimum Message Length, and Rissanen (1978)'s Minimum Description Length. Each of these procedures has its epistemic advantages, and some (e.g. Schwarz (1978)) have known asymptotic properties for a given class of likelihood functions.

In this paper, we suggest obtaining the required penalty by introducing priors on our parameters of interest (including  $k$ ), and finding the posterior mode estimates of  $(k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\})$ . Let  $X$  be the entire observed data set:  $(x_1^s, \dots, x_n^s)$  for  $s \in \{1, \dots, n\}$ , and let  $\Delta$  be the class of all matrices  $\{\delta_{ij}\}$ . Our posterior on the parameters of interest can be written as follows:

$$\Pr\{k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\} | X\} = \frac{1}{Denom} \Pr\{X | k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}\} \Pr\{k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}\}.$$

The second term on the right hand side is the probability of the data conditional on having  $k$  rules, those  $k$  rules being  $c^1, \dots, c^k$ , having error rate  $\epsilon$ , and allocating subjects to rules according to matrix  $\{\delta_{ij}\}$ . The third term is the joint prior on  $k$ , the rules  $(c^1, \dots, c^k)$ ,  $\epsilon$ , and the allocations of subjects to rules. *Denom* in the first term is defined by:

$$Denom = \int_{\epsilon \in [0,1]} \sum_{k=1}^{\infty} \sum_{\mathfrak{C}^k} \sum_{\Delta} \Pr\{X | k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}\} \Pr\{k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\}\} d\epsilon.$$

If we start with a prior  $\alpha_k$  on having  $k$  decision rules, and for  $k$  rules we introduce a prior  $\pi_k(c^1, \dots, c^k) \otimes \mu_k(d\epsilon) \otimes \nu_k(\{\delta_{ij}\})$ , we then calculate the posterior mode estimates  $(k, c^1, \dots, c^k, \epsilon, \{\delta_{ij}\})$ . We choose the prior  $\pi_k(\cdot)$  to be uninformative (assigning prior probability  $\frac{1}{83^k}$  to each possible  $k$ -tuple of rules in  $\mathfrak{C}^k$ ). (notice that this allows the same rule to be picked more than once in the  $k$ -tuple which may happen if there are in fact less than  $k$  rules in the population and we have free choice for the “unused” rules). We also choose the prior  $\nu_k(\cdot)$  to also be uninformative (assigning equal prior probability  $\frac{1}{k^n}$

to all possible allocations of the  $n$  subjects),  $\mu_k(d\epsilon)$  uniform for all  $k$ , and  $\alpha_k = \frac{1}{2^k}$ . Our posterior mode estimates are then obtained by maximizing:

$$\log\left(\prod_{s=1}^n \max_{h \in \{1, \dots, k\}} f^{\hat{c}^h, \hat{\epsilon}}(x_1^s, \dots, x_{t_s}^s)\right) - 3k \log(8) - k \log(2) - n \log(k).$$

Our procedure achieves consistent estimates (as the number of tasks that each individual performs in the experiment, and the number of individuals, go to infinity) of the number of rules being used, the rules themselves, and the proportion of the population using each of the rules (for proofs and technical details see El-Gamal and Grether (1993)).

We close this section by briefly comparing our approach to the vast and growing literature on classification and clustering. A primary goal of the classical classification literature is to establish simple algorithms that work for a large class of problems. For example, Wallace and Boulton (1968)'s Snob (and later Snob 2) program, and Cheeseman (1988)'s Autoclass II program assume normality of the data generating process. The general procedure we use agrees with all likelihood-based classification procedures in its form, but the class of likelihood functions is suggested by the problem. In that sense, we are closer to the coding theoretic approach to estimation and classification (e.g. Wallace and Freeman (1987), Rissanen (1987)).

## 5 The Data Analysis

The results that we obtained by applying our algorithm to the data from the four universities are reported in Table 1. We had a total of 257 subjects, and the total number of tasks was 4520 (the number of tasks per subject varied from one experimental session to another). Table 1 shows the estimated parameters for the four schools: UCLA, PCC, Occidental College, and CSULA, as well as estimates for the pooled sample of subjects who were payed according to the correctness of their guesses (All-pay), the pooled sample of subjects who were payed a flat fee (All-flat), and the pooled sample of all 257 subjects (All). For each  $k$ , we report the maximum likelihood estimates of  $\epsilon$ , the rules  $(c_1^1, c_2^1, c_3^1), \dots, (c_1^k, c_2^k, c_3^k)$ , the number of subjects allocated to each of the rules, the information criterion that we introduced in Section 4 ( $IC = \log(\text{maximal likelihood}) - 3k \log(8) - k \log(2) - n \log(k)$ ), a  $\chi^2$  goodness of fit statistic to be explained below, and the total number of data points for each subsample. When our information criterion told us to stop after a certain number of rules, we indicated that by using boldface for the estimates corresponding to that  $k$ . We now summarize the most important results:

1. For all tables but Occidental College and CSULA, when we force the algorithm to choose only one rule, it picks the rule (432), which corresponds to Bayesian updating. Even in the two institutions where (432) was not picked as the single most likely rule, when we allowed the algorithm to pick more rules, (432) surfaced. With the exception of PCC, the (432) Bayes rule has more subjects allocated to it than any other rule.

2. The second most prominent rule in all but the PCC table (where it is the most prominent) is (333). This is the “representativeness” rule, and its robustness regardless of the number of rules that we allow our algorithm to pick is also remarkable.
3. The third most prominent rule, once enough rules are allowed, is (531), a conservative rule. Our information criterion usually excludes this rule, but does not exclude weaker versions of conservatism that may be subsumed under (432).
4. Of our four schools, UCLA had the lowest estimated  $\epsilon$ , followed by PCC, Occidental College, and CSULA. Table 1 also shows that our estimate of  $\epsilon$  for the subjects who were payed according to the outcome was lower than its counterpart for the subjects who were payed a flat fee. A similar ordering is induced by the proportions of subjects who use Bayes’s rule.
5. In Table 2, we report, for  $k \in \{1, 2, 3, 4\}$ , the  $\chi^2$  statistics for likelihood ratio tests of homogeneity across schools and across payment schemes. The likelihoods for the pooled sample are estimated under the constraint that  $\epsilon$  and the rules used were invariant across all schools and payment schemes. We can therefore construct a likelihood ratio test for the null hypothesis that the value of  $\epsilon$  and the estimated rules are the same for both payment treatments by using the ratio of the likelihood in the pooled sample and the product of the likelihoods for the two payment schemes. Negative twice the likelihood ratio for each number of rules  $k$  is asymptotically distributed  $\chi^2$  with  $3k+1$  degrees of freedom. Similarly, we can take the ratio of the likelihood using the pooled sample to the product of the likelihoods estimated for each of the schools to construct a test statistic for the null of homogeneity across schools. The resulting statistic is asymptotically  $\chi^2$  with  $9k+3$  degrees of freedom. In Table 2, we report for each  $k$  the values of the  $\chi^2$  statistic for the two homogeneity tests (across schools and across payment schemes), together with the degrees of freedom, and the p-value of a test of homogeneity. The tests in Table 2 strongly reject the hypothesis that subjects in different schools are acting in similar ways, and strongly reject the hypothesis that subjects across different payment schemes act in similar ways. Other evidence on the effect of monetary incentives in similar contexts is equivocal, see e.g. Scott et al. (1988), and Wright and Aboul-Ezz (1988).
6. Our test of goodness of fit in Table 1 is a likelihood ratio test comparing the unconstrained estimated  $\epsilon$ ’s and rules, and a constrained model where we estimate  $\epsilon$  and a single rule for all agents in  $\{(-1, -1, -1), (6, 6, 6)\}$ . The rule  $(-1, -1, -1)$  corresponds to always choosing cage A, and the rule  $(6, 6, 6)$  corresponds to always choosing cage B, all other choices being explained by noise. The estimation of this model is very simple: count the number of A’s and B’s in each data set, assign all subjects to the rule that always makes the more common choice, and classify all other choices as errors. The likelihood is then

$$\mathcal{L}(A, B) = (1 - \min(A, B)/(A + B))^{\max(A, B)} \times (\min(A, B)/(A + B))^{\min(A, B)}.$$

This measure of goodness of fit is similar to a test of  $R^2 = 0$  in a binary choice model of the following kind:

$$y^* = \alpha + \beta X + u,$$

with  $y = 1$  if  $y^* > 0$ , and  $y = 0$  otherwise. The test of  $R^2 = 0$  in that model will be a test of how much of the variance of  $y^*$  is explained by the variance of  $u$ . In other words, we estimate  $\alpha$ , restricting  $\beta = 0$ , and compare the goodness of fit for  $\beta$  restricted and unrestricted to 0. Notice that in the restricted case of this model, if  $\hat{\alpha} > 0$ , then barring errors/noise  $u$ , the agent should systematically choose  $y = 1$ , and if  $\hat{\alpha} < 0$ , then if there is no noise, the agent will always choose  $y = 0$ . It is clear from Table 1 that we very strongly reject this model that explains the choices of  $A$  and  $B$  by noise, i.e. our estimated decision rules explain a significant proportion of the total variation in responses.

Some cautionary notes are in order. In finite samples, certain individuals' data could have the same likelihood under two or more rules, so the likelihood function will have multiple maxima. In Table 1, we gave the benefit of the doubt to rules that were a priori more appealing to us (e.g. (432) was chosen any time it was tied with one of the other rules). We simulated the percentage of time the data of an individual using the rule (432) can be tied with either (333) or (531) as a function of the number of tasks performed. We ran a Monte Carlo where, for number of tasks  $t$ , we drew 5000 artificial subjects using rule (432) and making errors with probability  $\epsilon = 0.3$ . It was clear from the Monte Carlo simulations that the probability of a tie declines to zero at a very fast rate. We then ran another Monte Carlo simulation using 5000 draws of artificial subjects who use the rule (432) and make errors with probability  $\epsilon = 0.3$ , and calculated the proportion of subjects using (432) being classified to either (333) or (531) under three tie-breaking rules. The first rule is the one we used in this paper and which always favors (432), the second rule uses 5000 random draws that do not have ties (i.e. ignores a draw that produces a tie and looks for another one), and the third rule randomly assigns a subject with a tie with equal probability (1/2 if a two-way tie, 1/3 if a three-way tie). The simulations show that the number of misclassifications also declines very quickly and that for a number of tasks close to the ones in our data, the proportion of misclassifications is around 15%.

It turns out in our application that the tie breaking rule was very seldom invoked. In the UCLA data with the three rules (432), (333) and (531) and  $\epsilon = 0.239$ , we invoked the rule for only 2 individuals out of 97 (both had equal likelihoods under (333) and (531)). In PCC, with three rules, we invoked the rule for only one of 56 subjects who could equally well be classified as (333) or (531). In Occidental College, when we restrict attention to 3 rules, we invoked the rule for 2 subjects out of 56 (both could be classified as either 333 or 511). On the whole, the tie breaking rule did not have much of an impact on our results.

On the positive side those problems arising in finite samples can be resolved not only by letting the number of tasks and the number of individuals get large, but also

by considerations of optimal experimental design. Our data set was not constructed for the purposes of this paper (see Grether (1980) for discussion). The data generating mechanism was chosen to increase the probability of getting outcomes (3 N's or 4 N's) that mimic the parent distributions of cages A and B. That design may be suboptimal for our purposes, and we probably would have chosen a larger number of priors, and more draws from the cages to reduce the probability of ties between various rules.

## 6 Comparison with the mixture model

So far, we have not discussed estimating the proportions of the different types (where a type is identified by the rule they use) in the population. Let  $\pi_j > 0; j = 1, \dots, k$ ; be the probability that each subject would be of type  $j$ , then we might want to estimate those probabilities from the estimated classifications (e.g.  $\hat{\pi}_j = \sum_{i=1}^n \hat{\delta}_{ij}/n$ ). Since the estimates of the  $\delta_{ij}$ 's are consistent, so will the estimates of the  $\pi_j$ 's.

In the mixture of types model, however, one can follow the literature (Redner and Walker (1984), Little and Rubin (1987)) by treating the  $\delta_{ij}$ 's as missing data, and employing the EM algorithm to estimate the expected values of the  $\delta_{ij}$ 's as well as the  $\pi_j$ 's and the rest of the parameters of the model. This procedure has some advantages over our estimation/classification procedure. First, since we classify each subject to the rule which maximizes their contribution to the likelihood function, we are in essence minimizing the number of errors attributed to each person. This results in a downward bias in our estimates of  $\epsilon$ . Moreover, the small sample misclassification errors are not taken into consideration when we estimate the rest of the parameters of the model.

The application of the EM algorithm to the mixture of types model would replace maximizing the likelihood function (4.1) with a maximization of expected log-likelihood. For the  $t^{\text{th}}$  iteration, given our current guesses  $p_{ij}^t$  for the expectations of the  $\{\delta_{ij}\}$ 's, the M-step produces the estimates:

$$(\hat{c}_t^1, \dots, \hat{c}_t^k, \hat{\epsilon}_t) = \operatorname{argmax}_{c^1, \dots, c^k, \epsilon} \sum_{s=1}^n \sum_{h=1}^k \hat{p}_{sh}^t \log \left( f^{c^h, \epsilon}(x_1^s, \dots, x_{t_s}^s) \right),$$

then, for the E-step, we define  $\hat{\pi}_j^t = (1/n) \sum_{i=1}^n \hat{p}_{ij}^t$ , and calculate  $\hat{p}_{ij}^{t+1}$  as the Bayes posterior with prior  $\hat{\pi}_j^t$ , and with the likelihood function calculated at  $(\hat{c}_t^1, \dots, \hat{c}_t^k, \hat{\epsilon}_t)$ . See Little and Rubin (1983) and Redner and Walker (1984) for the details of applying the EM algorithm in this framework.

It is well known (e.g. Dempster et al. (1977) and Tanner (1993, Chapter 4)) that the EM algorithm can be quite slow to converge, and performing the global optimization over all  $k$ -tuples of rules for each iteration of the algorithm (the M-step), the computational requirements of this procedure can be extremely high. However, it has the undeniable advantage of giving us natural estimates of the  $p_{ij}$ 's and the  $\pi_j$ 's as well as reducing the downward bias in our estimates of  $\epsilon$ . We note that if the  $p_{ij}$ 's are very close to

zeros and ones, the results from the EM algorithm and from our estimation/classification procedures would coincide. Moreover, with probability 1, as the number of tasks per individual goes to infinity, the two procedures will coincide and share the same consistency properties.

In Table 3, we fixed the estimated  $(\hat{c}^1, \dots, \hat{c}^k)$  of Table 1 for  $k = 2$  and 3, and ran the EM algorithm discussed above where at each M-step, we maximized only over  $\epsilon$ . The resulting estimates of the  $p_{ij}$ 's were very close to zeros and ones suggesting that our estimates of the rules are quite robust to the choice of estimation technique. The measure of divergence from zeros and ones that we show in this table is what we call average normalized entropy (ANE), which is simply  $-(1/n) \sum_{i=1}^n \sum_{j=1}^k p_{ij} \log_k(p_{ij})$ , where we use  $\log_k$  to make the maximal entropy equal to 1, and then we average the entropy over all individuals. If the  $p_{ij}$ 's were all zeros and ones, then ANE=0 would be diagnostic of very good behavior for the estimation/classification procedure, and a maximal ANE=1 would correspond to all the  $p_{ij}$ 's equal to  $(1/k)$ , with very poor performance of our procedure. The ANE's in Table 3 are all quite small; for instance,  $k = 2$  and  $\pi = (0.925, 0.075)$  or  $k = 3$  and  $\pi = (0.88, 0.1, 0.02)$  would produce an entropy of 0.38, the highest in Table 3. Moreover, the  $\hat{\epsilon}$ 's estimated from the EM algorithm are quite close to the ones that our procedure produced. We are, therefore, quite confident that the results provided by our simpler and less computationally demanding procedure are reliable for our data sets.

## 7 Concluding Remarks

The response of economists and psychologists to the discovery of anomalous violations of standard models of statistical decision theory has mainly been to devise new theories that can accommodate those apparent violations of rationality. The enterprise of finding out what experimental subjects actually do (instead of focusing on what they do not do; i.e. violations of standard theory) has not progressed to the point that one would hope. As a first step in that direction, we propose a general estimation/classification approach to studying experimental data. The procedure is sufficiently general that it can be applied to almost any problem. The only requirement is that the experimenter or scientist studying the experimental data can propose a class of decision rules (more generally likelihood functions) that the subjects are restricted to use. In many cases, such a class of rules may even be dictated by the experimental design itself. In El-Gamal and Grether (1993), we have shown that our proposed procedure has asymptotic optimality results which can be approximated in small samples by pre-selecting the experimental design to discriminate among the class of likelihood functions under consideration (for further discussions of optimal discrimination between a given class of models, see Boylan and El-Gamal (1993), El-Gamal et al. (1993), and El-Gamal and Palfrey (1994)).

Our first application of this procedure to experimental data dealing with decision making under uncertainty is (appropriately) targeted at the building block of any model of such decision making. Our results seem robust, and the most prominent rules that our algorithm selected are reasonable rules. The most prominent rule in most cases is

the Bayes updating rule. Hence, even though the answer to “are experimental subjects Bayesian?” is “no”, the answer to “what is the most likely rule that people use?” is “Bayes’s rule”. The second most prominent rule that people use is “representativeness”, which simply means that they ignore the prior induced by the experimenter, and make a decision based solely on the likelihood ratio. The third most prominent rule that our algorithm selects on the basis of the data is “conservatism”, which means that subjects give too much weight to the prior that is induced by the experimenter, needing more evidence to change their priors than the Bayes rule would imply. We believe that given the flexibility of our approach, and given the strong results that it generated in our particular application, its potential usefulness for uncovering the rules that are being used by experimental subjects can be quite substantial.

Table 1: Estimated rules, error rates, and classifications.

Sample	# Rules	Rule(s) chosen	#s class. to rule	$\hat{\epsilon}$	IC	$\chi^2$	N
UCLA	1	432	97	.308	-840.74	925.70	1940
PCC	1	432	67	.409	-482.35	328.54	938
Occ.Col.	1	332	56	.405	-479.84	332.16	939
CSULA	1	433	37	.484	-395.81	194.64	703
All-pay	1	432	125	.312	-1044.89	855.26	2123
All-flat	1	432	132	.457	-1147.83	1040.97	2397
All	1	432	257	.380	-2204.95	1862.85	4520
<b>UCLA</b>	<b>2</b>	<b>432,333</b>	<b>71,26</b>	<b>.261</b>	<b>-832.21</b>	<b>1091.09</b>	
<b>PCC</b>	<b>2</b>	<b>333,432</b>	<b>36,31</b>	<b>.277</b>	<b>-437.76</b>	<b>524.47</b>	
<b>Occ. Col.</b>	<b>2</b>	<b>332,511</b>	<b>47,9</b>	<b>.334</b>	<b>-476.56</b>	<b>430.23</b>	
<b>CSULA</b>	<b>2</b>	<b>432,333</b>	<b>26,11</b>	<b>.393</b>	<b>-387.66</b>	<b>276.10</b>	
<b>All-pay</b>	<b>2</b>	<b>432,333</b>	<b>85,40</b>	<b>.257</b>	<b>-1019.79</b>	<b>1092.62</b>	
<b>All-flat</b>	<b>2</b>	<b>432,333</b>	<b>77,55</b>	<b>.352</b>	<b>-1092.98</b>	<b>1346.13</b>	
<b>All</b>	<b>2</b>	<b>432,333</b>	<b>162,95</b>	<b>.302</b>	<b>-2108.88</b>	<b>2423.74</b>	
UCLA	3	432,333,531	50,26,21	.239	-837.60	1172.83	
PCC	3	333,432,531	36,25,6	.256	-453.13	561.92	
<b>Occ. Col.</b>	<b>3</b>	<b>432,333,511</b>	<b>28,19,9</b>	<b>.299</b>	<b>-474.57</b>	<b>493.48</b>	
CSULA	3	432,333,531	18,11,8	.370	-398.02	299.24	
All-pay	3	432,333,531	64,38,23	.234	-1022.30	1202.82	
All-flat	3	432,333,531	56,54,22	.325	1106.90	1438.88	
All	3	432,333,531	120,93,44	.277	-2121.91	2619.13	
UCLA	4	432,333,531,443	46,24,19,8	.229	-852.23	1213.25	
PCC	4	333,432,444,531	34,20,7,6	.239	-463.67	593.25	
Occ. Col.	4	432,332,333,511	21,14,13,8	.273	-479.28	530.13	
CSULA	4	432,333,421,542	15,11,7,4	.353	-406.53	317.39	
All-pay	4	432,333,531,433	55,31,22,17	.223	-1038.55	1256.11	
All-flat	4	333,432,521,542	52,51,19,10	.311	-1126.53	1488.36	
All	4	432,333,531,433	103,83,43,28	.267	-2160.00	2704.12	

Table 2: Tests of homogeneity.

No. of Rules	Across Schools			pay vs. flat fee		
	$\chi^2$	d.f.	p-value	$\chi^2$	d.f.	p-value
1	54.02	12	0.000	38.33	4	0.000
2	33.96	21	0.035	19.94	7	0.008
3	44.16	30	0.040	27.01	10	0.008
4	85.73	39	0.000	45.29	13	0.000

Table 3: EM algorithm estimates.

Sample	# Rules	Rule(s) chosen	$\pi_j$ for each rule	$\hat{\epsilon}$	ANE
UCLA	2	432,333	.79 , .21	.270	.243
PCC	2	333,432	.53 , .47	.290	.362
Occ. Col.	2	332,511	.84 , .14	.340	.111
CSULA	2	432,333	.66 , .34	.400	.313
All-pay	2	432,333	.74 , .26	.302	.326
All-flat	2	432,333	.59 , .41	.322	.356
All	2	432,333	.66 , .34	.312	.343
UCLA	3	432,333,531	.61 , .22 , .17	.250	.364
PCC	3	333,432,531	.54 , .40 , .06	.272	.274
Occ. Col.	3	432,333,511	.55 , .32 , .13	.308	.324
CSULA	3	432,333,531	.66 , .34 , .00	.392	.384
All-pay	3	432,333,531	.61 , .26 , .13	.202	.343
All-flat	3	432,333,531	.47 , .41 , .12	.304	.352
All	3	432,333,531	.54 , .34 , .12	.296	.350

Table 4: Entries are #A's and #B's for each prior and outcome aggregated over the eight experimental sessions.

#N's	0	1	2	3	4	5	6
prior = 1/3	1 , 47	8 , 69	10 , 111	39 , 405	263 , 424	157 , 51	0 , 0
prior = 1/2	0 , 0	3 , 29	38 , 267	22 , 138	292 , 79	181 , 20	42 , 9
prior = 2/3	0 , 0	15 , 71	61 , 210	192 , 146	613 , 69	256 , 12	157 , 13

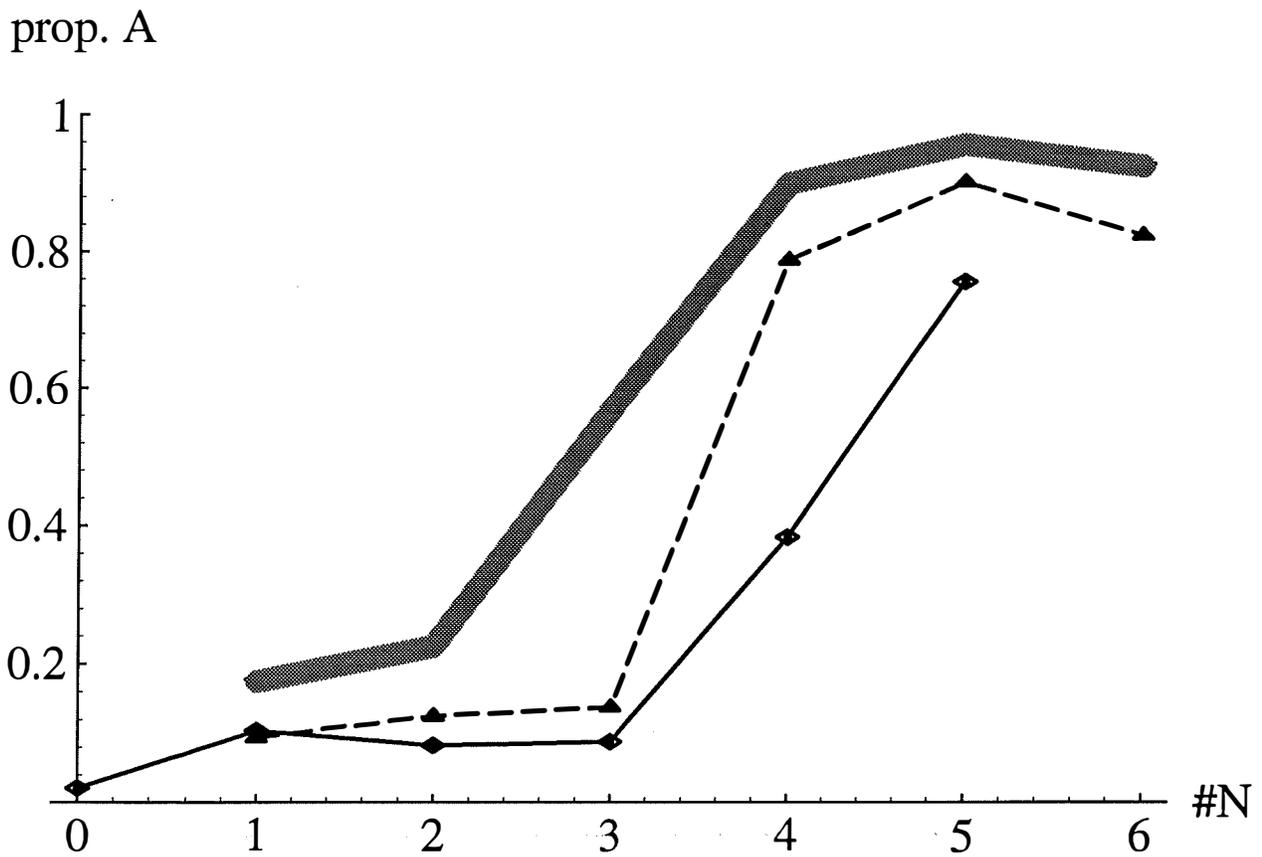


Figure 1: Proportion of A choices as a function of number of N's observed for priors 1/3 (solid black line), 1/2 (dashed line), and 2/3 (solid gray line).

## References

- Akaike, H. 1974. A new look at the statistical identification model. *IEEE Transactions on Automatic Control* 19:716–723.
- Allais, M. 1953. Le comportement de l’homme rationel devant le risque, critique des postulats et axiomes de l’école américaine. *Econometrica* 21:503–546.
- Bell, D. 1982. Regret in decision making under uncertainty. *Operations Research* 30:961–981.
- Birnbaum, M., G. Coffey, B. Mellers, and R. Weiss. 1992. Utility measurement: configural-weight theory and the judge’s point of view. *Journal of Experimental Psychology: Human Perception and Performance* 18(2):331–346.
- Bostic, R., R. Herrnstein, and D. Luce. 1990. The effect on the preference-reversal phenomenon of using choice indifference. *Journal of Economic Behavior and Organization* 13:193–212.
- Boylan, R. and M. El-Gamal. 1993. Fictitious play: A statistical study of multiple economic experiments. *Games and Economic Behavior* 5(2):205–222.
- Cheeseman, P. 1988. Autoclass II conceptual clustering system. *Proceedings of Machine Learning Conference* pages 54–64.
- Chew, S. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the allais paradox. *Econometrica* 51:1065–1092.
- Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38.
- Edwards, W. 1982. Conservatism in human information processing. In Kahneman, D., P. Slovic, and A. Tversky, eds., *Judgment Under Uncertainty: Heuristic and Biases*, pages 359–369. Cambridge University Press, Cambridge.
- El-Gamal, M. and D. Grether. 1993. Uncovering behavioral strategies: Likelihood-based experimental data mining. Soc. Sc. Working Paper #850. Caltech.
- El-Gamal, M., R. McKelvey, and T. Palfrey. 1993. A Bayesian sequential experimental study of learning in games. *Journal of the American Statistical Association* 88:428–435.
- El-Gamal, M. and T. Palfrey. 1994. Economical experiments: Bayesian efficient experimental design. Soc. Sc. Working Paper #884. Caltech.
- Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75:643–669.

- Gigerenzer, G., H. Ulrich, and H. Kleinbolting. 1991. Probabilistic mental models: A brunswikian theory of confidence. *psychological Review* 98(4):506–528.
- Goldstein, W. and H. Einhorn. 1987. Expression theory and the preference reversal phenomena. *Psychological Review* 94(2):236–254.
- Grether, D. 1980. Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics* pages 537–557.
- Grether, D. and C. Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69:623–638.
- Kahneman, D. and A. Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3:430–454.
- Kahneman, D. and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–292.
- Lichtenstein, S. and P. Slovic. 1971. Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89:46–55.
- Little, R. and D. Rubin. 1983. On jointly estimating parameters and missing data. *The American Statistician* 37:218–220.
- Little, R. and D. Rubin. 1987. *Statistical analysis with missing data*. New York: Wiley.
- Loomes, G., C. Starmer, and R. Sugden. 1989. Preference reversal: information-processing effect or rational non-transitive choice? *Economic Journal* 99:140–151.
- Loomes, G. and R. Sugden. 1987. Some implications of a more general form of regret theory. *Journal of Economic Theory* 41:270–287.
- Machina, M. 1982. Expected utility analysis without the independence axiom. *Econometrica* 50:277–323.
- Mellers, B., L. Ordonez, and M. Birnbaum. 1992. A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes* 52:331–369.
- Payne, J. 1982. Contingent decision behavior. *Psychological Bulletin* 92(2):382–402.
- Quigen, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3:323–343.
- Redner, R. and H. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. *Siam Review* 26, no.2:195–239.
- Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.
- Rissanen, J. 1987. Stochastic complexity. *Journal of the Royal Statistical Society B* 49(3):223–239 and 252–265.

- Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6,2:461–464.
- Scott, W., J. Farg, and P. Podsakoff. 1988. The effect of ‘intrinsic’ and ‘extrinsic’ reinforcement contingencies on task behavior. *Organizational Behavior and Human Decision Processes* 41:405–425.
- Tanner, M. 1993. *Tools for statistical inference*. New York: Springer-Verlag.
- Tversky, A. and D. Kahneman. 1972. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 51:207–232.
- Tversky, A. and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131.
- Tversky, A., S. Sattath, and P. Slovic. 1988. Contingent weighting in judgement and choice. *Psychological Review* 95(3):371–384.
- Wallace, C. and D. Boulton. 1968. An information measure for classification. *computer Journal* 11:185–195.
- Wallace, C. and P. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B* 49(3):240–265.
- Wright, W. and M. Aboul-Ezz. 1988. Effects of extrinsic incentives on the quality of frequency assessments. *Organizational Behavior and Human Decision Processes* 41:143–152.
- Yaari, M. 1987. The dual theory of choice under risk. *Econometrica* 55(1):95–116.