On ignoring scientific evidence: The bumpy road to enlightenment

Robin M. Hogarth*

ICREA & Universitat Pompeu Fabra, Barcelona

robin.hogarth@upf.edu

May 19, 2006

* Robin M. Hogarth is ICREA Research Professor at Universitat Pompeu Fabra, Barcelona, Spain. He is grateful for comments on an earlier version of this work by Robyn M. Dawes, Spyros Makridakis, and J. Scott Armstrong. This research was financed partially by a grant from the Spanish Ministerio de Educación y Ciencia.

Abstract

It is well accepted that people resist evidence that contradicts their beliefs. Moreover, despite their training, many scientists reject results that are inconsistent with their theories. This phenomenon is discussed in relation to the field of judgment and decision making by describing four case studies. These concern findings that "clinical" judgment is less predictive than actuarial models; simple methods have proven superior to more "theoretically correct" methods in times series forecasting; equal weighting of variables is often more accurate than using differential weights; and decisions can sometimes be improved by discarding relevant information. All findings relate to the apparently difficult-to-accept idea that simple models can predict complex phenomena better than complex ones. It is true that there is a scientific market place for ideas. However, like its economic counterpart, it is subject to inefficiencies (e.g., thinness, asymmetric information, and speculative bubbles). Unfortunately, the market is only "correct" in the long-run. The road to enlightenment is bumpy.

The concept of mental models is a useful way of thinking about how people make sense of and understand what happens in the world. Thus, in dealing with the physical world, humans share similar mental models concerning the effects of gravity. For example, if you let something slip from your hands, you expect it to fall. Many of our models are the result of our interactions with the world and are largely tacit in nature (Hogarth, 2001). On the other hand, models can also be formalized and communicated explicitly to others. Indeed, this is one way of describing what scientists do.

What happens, however, when events in the world do not conform to the predictions (implicit or explicit) of your model? Imagine, for example, that when you let something slip out of your hand, it floats instead of falling. Do you question your eyesight or your model? Or do you ask whether you are in strange conditions where the model "does not apply"? Note that this, essentially, is what scientists should do when they first meet surprising phenomena (where surprising means relative to model-based expectations).

Surprising results can have three causes: (1) the method used to obtain the result was flawed (in the example just given, perhaps there is something wrong with your eyesight?); (2) the model really is incorrect (left by themselves, objects do float instead of fall); and (3) there are specific circumstances – perhaps not previously encountered – where the model does not apply (perhaps you observed the object while traveling in a space vehicle where gravity has no effect?).

Relative to our ability to understand, there is no question that the world is complicated. Thus, the models (and theories) we hold represent the accumulation of both our own experience and that of our ancestors. This knowledge – although

imperfect – has taken considerable time to develop and thus, when the predictions of models fail, one can understand why people do not wish to abandon cherished beliefs.

When a theory fails, what should one do? First, it is appropriate to question the methodology that yielded the erroneous prediction. Clearly, one should discount results produced by inappropriate methods. However, what if the methodology is appropriate and, in addition, several replications confirm the original results? If this is the case, it seems almost trivial to state that the model should be amended – either rejected as incorrect or specified to be more limited than originally thought. However, the history of science is replete with examples where this does not happen. Indeed, some time ago Kuhn (1962) brilliantly described the difficulty of replacing obsolete scientific paradigms (see also below).

The purpose of this chapter is to discuss this phenomenon with respect to the field of judgment and decision making. There are two reasons why this field provides an interesting setting for this issue. First, for scientists concerned with how decisions are and should be made, one might imagine that there would be little resistance to adopting methods that improve decision making by increasing accuracy, simplifying use, or both. Second, the studies in which these new results were discovered are empirical and often supported by analytical rationales. A priori, it is not a question of dubious evidence.

The chapter will discuss four cases of this phenomenon. These are, first, the findings that predictions of "clinical" judgment are inferior to actuarial models; second, how simple methods in times series forecasting have proven superior to more sophisticated and "theoretically correct" methods advocated by statisticians; third, how in combining information for prediction, equal weighting of variables is often more accurate than trying to estimate differential weights; and fourth, the observation that, on occasion, decisions can be improved when relevant information is deliberately

discarded. As a general statement, one theme underlies all four cases. This is that simple models can perform "better" than more complicated ones. However, this is a difficult principle for people to grasp. When making decisions perceived as complex, there is a strong belief that our methods or use of information should match the complexity of the situation. Before presenting the four cases, I first comment briefly on how – based on the psychological literature – one might expect people to react to evidence that disconfirms their theories as well as discussing several notable cases from the history of science.

Do scientists revise their theories in light of new evidence?

The Bayesian model provides a way of thinking about how people should revise beliefs in the light of new evidence. The model essentially suggests a three-step process: (1) a state of belief that a particular theory is "correct" or the "best" at time t_1; (2) the arrival of new evidence at time t_2 that is evaluated as being favorable or disfavorable to the theory; and (3) the incorporation of (2) with (1) at time t_3 such that a new, revised level of belief is reached that takes account of both the prior level (1) and the direction and strength of the evidence (2). Abstracting from technical difficulties of applying the Bayesian model in practice, we can nonetheless emphasize several important qualitative implications.

First, if somebody has a dogmatic belief that a theory is correct or incorrect (p = 1,0), then no evidence can change this. Second, assessment of the direction and strength of the evidence in (2) should be independent of the level of the beliefs held in (1). Third, provided people are not dogmatic, i.e., $0 < p < 1$ in (1), by observing the same sequence of evidence across time they will eventually converge on the same estimate of p. How long this takes, of course, depends on the dispersion of initial

beliefs and the evidential strength of what is observed. Let us consider some evidence on each of these implications.

Whereas we can use the Bayesian model to examine beliefs in specific propositions, it is important to emphasize that any particular belief is likely to be part of a system of inter-connected beliefs that a person holds when thinking about the world. For example, the famous physicist Lord Kelvin (who also became infamous for his refusal to accept several important, scientific discoveries) claimed that he could never understand a phenomenon unless he could make a mechanical model of it and, for this reason, could not "get the electromagnetic theory;" nor did Kelvin ever abandon "the concept that the atom is an indivisible unit" (Barber, 1961, p. 598). In short, it can be difficult for people to view scientific beliefs in "isolation." Beliefs are interconnected. At the same time, this does not mean that people are incapable of separating beliefs or that all belief systems are coherent. As casual empiricism demonstrates, many people hold beliefs simultaneously in different propositions that are mutually contradictory.

A second issue relates to the precision of beliefs. In many cases beliefs are not precisely formulated nor are they held with precise degrees of belief. For example, consider a widespread belief among researchers in judgment and decision making that people are "over-confident" (see, e.g., Bazerman, 1997). What does this really mean? That people always express more confidence in their judgments than justified by subsequent events? That people are sometimes overconfident? That some people (who might or might not be defined) are overconfident in some situations? And so on. In other words, because the general belief is not stated in precise, operational terms it is unclear how it should be affected by subsequent evidence.

Let us now consider the evaluation of evidence where, in particular, selective attention to either parts of the evidence or how the evidence was produced can affect

how much someone allows evidence to affect their beliefs. The "culprit" here is that evidence is not assessed independently of prior beliefs.

A classic study on this phenomenon was conducted by Lord, Ross, and Lepper (1979) who demonstrated, in a social psychology experiment, how the beliefs of participants (for or against capital punishment) were strengthened after they heard evidence contrary to their beliefs (see also Festinger, Riecken & Schachter, 1956). In reviewing literature on this topic in science, Koehler (1993) stated:

> Ian Mitroff (1983) conducted a series of detailed interviews with 42 eminent Apollo moon scientists and reported that most were emotionally involved in their work. Furthermore, those who held very strong beliefs about the nature of the moon appeared most anxious to dismiss evidence that contradicted their personal theories. Similarly, Mahoney (1977) studied a group of 75 scientific journal reviewers and found that they were strongly biased against manuscripts that reported results contrary to their strong behaviorist perspective. In short, judgments about the quality of scientific research appear to be quite dependent on the fit between a scientist's own beliefs and the conclusions supported by the research, particularly when the beliefs are strongly held. (Koehler, 1993, pp 29-30).

In his own research, Koehler (1993) demonstrated how alliance with particular scientific theories affects judgments of quality of studies and thus their potential evidential impact. Not only did Koehler study two groups of practicing scientists on opposite sides of a particular issue. Of greater concern is the fact that he also demonstrated that these effects can occur when graduate students in the physical sciences are endowed (at random) with opposing theories.

In short, there is ample evidence that people – including scientists – violate the qualitative implications of the Bayesian model of belief updating. And, as Armstrong (1997) has documented, the peer review process for evaluating scientific work does not alleviate and may even exacerbate these dysfunctional tendencies.

In a fascinating review, Barber (1961) has documented many cases involving scientific giants operating in the physical sciences where, one might suppose, hard evidence would be difficult to overcome. Among the various social influences or resistance to new ideas, Barber gives examples due to difficulties of understanding substantive concepts, different methodological conceptions, religious ideas, professional standing (e.g., failure to accept discoveries by young scientists), professional specialization (e.g., work by people outside a discipline), and the dysfunctional role sometimes played by professional societies. He goes on to quote Max Planck who, frustrated by the fact that his own ideas were not always accepted, stated

> A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it. (Barber, 1961, p. 597).

Four case studies

1. Clinical versus statistical prediction. A book published by Paul Meehl in 1954 is the first case I consider. In this book, Meehl asked the question whether – in predictions made in clinical psychology – clinicians would be better off using statistical aggregations of the data available on clients or alternatively to rely on their traditional method of unaided clinical judgments, i.e., subjective interpretations based on all data available to them. Meehl reviewed some 20 studies and discovered, provocatively, that the statistical method of prediction was superior to what is known as the "clinical" method.

At one level, one might have thought that this finding would have been welcome. After all, clinical prediction is both time consuming and important and, if a method could be devised that was both cheaper and more accurate, surely this would be adopted as being in everyone's interest. Nothing could be further from the case.

Clinicians were outraged by the implications of Meehl's study. The use of statistical formulas instead of trained professionals was seen as degrading. The study also struck at the heart of an important debate in the philosophy underlying clinical psychology, namely the extent to which the science should be nomothetic (concerned with general laws that apply to groups of people) or idiographic (concerned with particular individuals). Many clinicians who found Meehl's results distasteful were clearly in the latter group (Holt, 1962).

The most eloquent – and persistent – of Meehl's critics has been Holt (1958; 2004). It is therefore instructive to consider the kinds of arguments that were brought to bear against Meehl's findings. In Holt (1958), we find several attempts to suggest that comparing clinical and statistical judgment in the manner done by Meehl (1954) was just inappropriate. Thus Holt states,

> …clinicians do have a kind of justified grievance against Meehl, growing out of his formulation of the issues rather than his arguments, which are sound (p. 1).

Meehl's comparisons, it is claimed, were unfair to the clinicians because, unlike clinical judgments, actuarial predictions had been cross-validated, thus:

> ….in none of the studies Meehl cites were the clinical predictions under test being cross-validated. This alone is a major reason to expect superior performance from the actuarial predictions, and again it is a disadvantage under which the clinician by no means has to labor (p. 3).

Holt goes on to argue that the process of clinical prediction involves various phases and that Meehl's comparisons did not match like with like and thus "in none of the 20 studies Meehl cites were the comparisons pertinent to the point" (p. 4). In other words, Holt rejected both the problem, as formulated by Meehl, as well as the specific comparisons he made as being irrelevant. He also went on to suggest a conceptual

framework for prediction that he claimed was more "scientific" than the studies reviewed by Meehl.

What is curious is that although Holt's article contains many good points about aspects of the clinical process where human judgment is essential, he never wants to accept that there are situations where clinical "intuition" might tradeoff with the consistent use of decision rules (cf., Goldberg, 1970). Also, it is clear that there are problems for which it is infeasible to build adequate statistical models and where clinical judgment is necessarily better than actuarial formulas (see, e.g., Yaniv & Hogarth, 1993). Indeed, from Garb's (1998) comprehensive review, it is clear that clinical judgments are far from being universally ineffective in a relative sense.

In the half century that followed the publication of Meehl's book, many studies have reinforced the original findings (see, e.g., Sawyer, 1966; Dawes, Faust, & Meehl, 1989; Kleinmuntz, 1990). In 2000, a meta-analysis by Grove, Zald, Lebow, Sniz, and Nelson summarized the results of 136 studies comparing clinical and statistical judgments across a wide range of task environments. Their findings did not show that statistical methods were always better and, in fact, they identified a few studies in which clinical judgment was superior. On the other hand, they summarized their results by stating

> ….we identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges (Grove et al., 2000, p. 25).

As evident from this meta-analysis, it is clear that the implications of Meehl's original insights go beyond the clinical-statistical debate in psychology and apply to any area of activity where data need to be aggregated in a consistent manner. Computers are just much better at this task than humans and yet, depending on the kind of task that is

considered, people have difficulty in accepting this fact. Let me illustrate by one further study and some observations.

In 1972 Hillel Einhorn published a study of judgments made by physicians who were experts in a certain form of cancer. The physicians' task was to view biopsy slides taken from patients and to (a) define the level of presence/absence of different indicators of disease in the slides and (b) estimate the overall severity of the disease as evidenced by the slides. Einhorn used the study to demonstrate the combined effectiveness of humans and computers as opposed to the use of humans or computers alone. He did this by showing that a statistical model that aggregated the physicians' judgments of cues, (a) (i.e., levels of indicators), was a more effective predictor of outcomes (length of patients' survival) than the physicians' judgments alone, i.e., (b). Einhorn's point was that better outcomes could be achieved by a system of "expert measurement and mechanical combination" than by a system that only relied on the expert physicians. In this particular case, the physicians' judgments of (a) were essential to the development of the model because there was no other way of measuring these cues. Einhorn's point was not to denigrate the expertise shown by the physicians in their reading of the biopsy slides. However, the physicians felt quite clearly that the study was an unfair condemnation of their abilities and became quite defensive about it.

Parenthetically, by a peculiar twist of fate, Einhorn in fact suffered from the same disease that the physicians were attempting to predict. Subsequently, I used the same dataset in my PhD thesis (Hogarth, 1972; 1974). When I attempted to contact the physicians with questions, the initial reaction was that I should not be allowed to use the data.

My second observation arises from an experience involving a large academic program. Here, the director of admissions spent an enormous amount of time each year

reading applications before using "clinical" judgment to make decisions. A faculty committee studied the admissions process and suggested using a statistical model based on the information in the application files. The suggestion was not well received even though it was stated that the model should only be used to pick the top 10% for admission and to reject the lowest 10% (thereby economizing some 20% of application reading time). The director clearly felt that the model was an intrusion into his domain of expertise (see also Dawes, 1979). Moreover, it would no longer allow him to claim that he read all files personally.

On the other hand, there are situations where the clinical-statistical controversy is well understood and has huge economic consequences. Consider, for example, the use of credit-scoring by banks and finance companies. For many kinds of accounts, these corporations no longer rely on "clinical" procedures when granting credit. Instead, they rely on simple models with a handful of variables (sometimes as few as 1 or 2) to predict which potential clients are or are not good credit risks. (For an interesting application, see Showers & Chakrin, 1981). In these applications, economic incentives certainly seem to make a difference.

2. Simple models in time series. A critical operational concern in economics and business (private and public) is the forecasting of many different time series. Consider, for example, data concerning imports and exports across time, the supply and demand for specific products and classes of goods, inventories, and various economic indicators. Forecasting these variables with a reasonable level of accuracy is essential because, without good forecasts, individuals and firms cannot plan and economic activity suffers.

Since the 1950s and 1960s the availability of computers has considerably increased the ability to forecast millions of time series. At the same time, theoretical statisticians have spent considerable effort developing increasingly sophisticated

methods for determining patterns in time series with the ostensible objective of achieving better predictions.

However, it was not until the 1970s that statisticians first started to question which particular methods might work better for predicting actual series in practice. These first studies (e.g., Newbold & Granger, 1974) compared comparatively few methods (see below) and, although their results were not unambiguous, were generally supportive of the status quo models in the theoretical statistical literature (Box & Jenkins, 1976).

In 1979, Spyros Makridakis and Michèle Hibon (at the time comparatively unknown researchers) broke with tradition by presenting a paper at the prestigious Royal Statistical Society in which they compared the out-of-sample forecasting performance of 22 forecasting methods on 111 time series they had obtained from various sources in business and economics. Their methodology was conceptually simple: separate each time series into a fitting phase and predictive phase; fit all models on the fitting data; use the fitted models to make predictions for the predictive phase; and compare predictions with realizations (i.e., similar to cross-validation in using multiple regression).

Results surprised even the authors: "…if a single user had to forecast for all 111 series, he would have achieved the best results by using exponential smoothing methods after adjusting the data for seasonality" (Makridakis & Hibon, 1979, p. 101). In other words, a very simple model (that essentially only weights the last few observations) outpredicted many complex and statistically sophisticated models that provided closer fits to the data in the fitting phase of the analyses. The essential point made by Makridakis and Hibon was also conceptually simple: real-world time series in business and economics are not necessarily stationary (in the statistical sense) and thus extreme

caution should be observed in predicting out-of-sample. Complicated models may not be worth the cost.

Comments made at the meeting, and afterwards, were published by the *Journal of the Royal Statistical Society* and make interesting reading today. Between the compliments for conducting a demanding empirical study and legitimate questions about methodology, there were several published statements that were clearly intended to dismiss the results. For example, one prominent commentator stated:

> *If the series conforms to an ARMA model, and the model has been fitted correctly, then the forecast based on this ARMA model must, by definition, be optimal.* (Apart from the ARMA model, all the other forecasting methods considered are of an *ad hoc* nature. The ARMA method involves model fitting and its performance depends to a large extent on the ability of the user to identify correctly the underlying model.) (Italics and parentheses in original, Priestley, 1979, p. 128).

As noted, the commentator did not appear to want to be concerned by empirical evidence and also hinted that the investigators had not followed appropriate procedures. Other commentators wondered whether there was something peculiar about the particular time series the authors had assembled. One went so far as to state that Makridakis's competence to perform appropriate time-series analyses should not be trusted.

Makridakis's reactions since 1979 have been exemplary. In 1982, he published results of the M-competition (Makridakis et al., 1982) in which experts in different forecasting methods were invited to predict 1001 series (thereby avoiding the criticism that he had used the methods inappropriately). In 1993, results of the M-2 competition were made available (Makridakis et al., 1993). This competition was similar to the M competition in that experts were invited to use their own methods. It differed, however, in that there were fewer forecasts but these were conducted in real time (e.g., you are asked now to provide a forecast for next year). Moreover, forecasters could obtain

background and qualitative data on the series they were asked to forecast (a criticism of the M competition was that experts lacked access to important contextual information). Finally, results of the M-3 competition appeared in Makridakis and Hibon (2000). In this, forecasts were prepared for several models using 3003 time series drawn from various areas of economic activity and for different forecast horizons. In addition to these M-competitions, other scholars have conducted their own similar studies and essentially replicated the earlier findings of Makridakis and Hibon, namely:

> (a) Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones. (b) The relative ranking of the performance of the various methods varies according to the accuracy measure being used. (c) The accuracy when various methods are being combined outperforms, on average, the individual methods being combined and does very well in comparison to the other methods. (d) The accuracy of the various methods depends on the length of the forecasting horizon involved. (Makridakis & Hibon, 2000, p. 452.)

One might imagine that, with this weight of evidence, the academic forecasting community would have taken on the important task of explicating the empirical evidence and developing models that could explain the interaction between model performance and task characteristics. However, there seems to be little evidence of this occurring. For example, Fildes and Makridakis (1995) used citation analysis in statistical journals to assess the impact of empirical forecasting studies on theoretical work in time-series analysis. Basically, their question was whether the consistent out-of-sample performance of simple forecasting models had led to theoretical work on illuminating this phenomenon. The answer was a resounding "no."

> Empirical validation, comparative modeling and the choice between alternative models (and methods) seem to have been regarded as unimportant by theoreticians in the field of statistical forecasting. …….the evidence is straightforward: those interested in applying forecasting regard the empirical studies as directly relevant to both their research and to applications…….those interested in developing statistical models….pay little attention or ignore such studies. (Fildes & Makridakis, 1995, p. 300).

Since this study had been published in 1995, I contacted Makridakis as to whether the situation had changed in the interim. The answer was in the negative (Spyros Makridakis, personal communication, January 2005).

Once again, it seems that whereas direct economic incentives have an important impact, scientists do not necessarily see the implications of negative evidence.

3. Unit or equal weighting – the power of averaging. During their studies, most social scientists learn the statistical technique of multiple regression. Given observations on a dependent variable y_i ($i = 1,\ldots, n$) and k independent or predictor variables x_{ij} ($j = 1,\ldots,k$), the budding scientists learn that the "best" predictive equation for y expressed as a linear function of the x's is obtained by the well-known least squares algorithm. The use of this technique (and more complex adaptations of it) is probably most used for hypothesis testing. Is the overall relationship statistically significant (i.e., is $R^2 > 0$?). What are the signs and relative sizes of the different regression coefficients? Which are more important? And so on.

In addition to describing data, another important function of multiple regression is to make predictions. Given a new (hold-out) sample of x's, what are the associated predicted y values? In using a regression equation in this manner, most users understand that the R^2 achieved on initial fit of the model will not be matched in the predictive sample due to "shrinkage" (the smaller the ratio n/k the greater the shrinkage). However, they do not question that the regression weights initially calculated on the "fitting sample" are the best that could have been obtained and thus that this is still the optimal method of prediction.

In 1974, Dawes and Corrigan reported the following interesting experiment. Instead of using weights in a linear model that have been determined by the least squares algorithm, use weights that are chosen at random (between 0 and 1) but subject

to having the appropriate sign. The results of this experiment were most surprising to scientists brought up in the tradition of least-squares. The predictions of the quasi-random linear models were quite good and, in fact, on four datasets, they exceeded the predictions made by human judges who had been provided with the same data (i.e., values of the predictor variables). This result, however, did not impress referees at the *Psychological Review* who rejected the paper. It was deemed "premature." In addition, the authors were told that, despite their results, differential regression coefficients are important for describing the strategies of judges. Subsequently, and before the paper appeared in the *Psychological Bulletin*, Dawes presented the results at a TIMS/ORSA conference only to be told by distinguished attendees that the results were "impossible." On the other hand, it should be added that some scientists who had heard one of Dawes's earlier talks on this subject tried out the "method" on their own datasets and saw that it worked (Robyn Dawes, personal communication, December 2004).

Dawes and Corrigan outlined four reasons for the success of their method: (1) in prediction, having the appropriate variables in the equation may be more important than the precise form of the function; (2) each predictor has a conditionally monotone relationship with the criterion; (3) the presence of error of measurement; and (4) deviations from optimal weighting may not make much practical difference. Subsequently, Einhorn and I examined the phenomenon analytically (Einhorn & Hogarth, 1975).

To do so, we first transformed the Dawes and Corrigan model by assuming an equal weight model (i.e., all regression coefficients are given equal weight) subject only to knowing the correct sign (zero-order correlation) of each variable. (This is the same as Dawes and Corrigan's model if one uses the expected values of the random weights.) We then went on to show the rather general conditions under which such equal- or unit-

weighting models correlate highly with so-called optimal weights calculated using least squares. Furthermore, we indicated how predictions based on unit weights are not subject to shrinkage on cross-validation and that conditions exist under which such simpler models would predict more accurately than ordinary least squares. In fact, prior to the appearance of both our paper and that of Dawes and Corrigan, several other papers had hinted at these results (see, in particular, Wilks, 1938; Claudy, 1972; Schmidt, 1971). In addition, Wainer (1976) published an article in the *Psychological Bulletin* with the catchy title "Estimating coefficients in linear models: It don't make no nevermind" in which he also showed that least-squares regression weights could often be replaced by equal weights with little or no loss in accuracy.

By this time, with both empirical and analytical results available, one might imagine that users of regression techniques would now know that sets of regression coefficients cannot be interpreted unambiguously. Moreover, to show real effects of differential sizes of coefficients, one should put estimated models to predictive tests where equal weight models provide a baseline. However, it is hard to find examples of this level of understanding in the literature. It is not true to say that the original papers have been ignored. Indeed, on February 24, 2005, the ISI Web of Knowledge reported that Dawes and Corrigan (1974) had been cited 663 times. Moreover, a number of studies in the decision making literature have exploited the results. However, the implications of this work have had surprisingly little impact on the methods of scientists who make great use of regression analysis.

Economists, for example, are among the most sophisticated users of regression analysis. I therefore sampled five standard textbooks in econometrics to assess what young economists are taught about ambiguity in regression weights and whether the benchmarks of equal or unit-weighting models for prediction were explained. The

specific textbooks were by Greene (1991), Griffiths, Hull, and Judge (1993), Goldberger (1991), Johnston (1991), and Mittelhammer, Judge, and Miller (2000). The answer was an overwhelming "no." The major concern of the texts seems to lie in justifying parameter estimates through appropriate optimization procedures. The topic of prediction using regression is given little space, and when it is, emphasis is placed on justifying regression coefficients in the prediction equations that have been estimated on the data available. None of the books gives any attention to equal- or unit-weighting; nor are any references made to the work of Dawes (let alone of Einhorn and Hogarth). My hopes rose when I located a chapter on "Evaluating the predictive accuracy of models" in a handbook whose contributors were leading econometricians. However, the chapter on this topic by Fair (1986) showed no awareness of the equal-weight findings. Paradoxically, I remember meeting Fair in France in the 1970's and telling him about the equal weights results. Our conversation clearly had no impact.

In psychology, on the other hand, the development of test theory has meant that students are implicitly instructed in the properties and use of equally-weighted composite variables (cf., Ghiselli, Campbell, & Zedeck, 1981). Indeed, the 3rd edition of Nunnally and Bernstein's *Psychometric Theory* (1994) explicitly devotes a section of a chapter (p. 154) to equal weighting citing, among others, Dawes and Corrigan (1974) and Wainer (1976). Interestingly, they emphasize that using equal weights is important when questions center on prediction in applied problems.

How does one explain the relative lack of interest in equal weights? In particular, contrary to the two cases examined above where evidence was restricted to empirical results, the case against naively accepting estimates of regression coefficients has been made on both empirical and analytical grounds. Perhaps, the reason is that there is a huge "industry" propagating the use of regression analysis involving

textbooks, computer software, and willing consumers who accept analytical results with little critical spirit, somewhat similar in manner to the use of significance tests in reports of psychological experiments (cf., Gigerenzer, 1998). Just because ideas are "good," does not mean that they will be presented in textbooks and handed down to succeeding generations of scientists (see, for example, the discussion by Dhami, Hertwig, & Hoffrage, 2004 concerning the concept of representative design).

Finally, above I referred to this case as involving "the power of averaging" because, in effect, the equal weighting model correlates perfectly with the arithmetic mean of the *x* variables (assuming that they have equal standard deviations). Curiously, people have poor intuitions about how useful the average can be when aggregating data. For example, as noted above by Makridakis and Hibon (2000), the average of several forecasts is typically one of the more accurate of the forecasts averaged (see also Hogarth, 1978). Indeed, some time ago social psychologists discovered that to guess a quantity (e.g., the number of jelly beans in a jar), one of the best methods was simply to average the estimates of different individuals (Gordon, 1924). Similarly, Larrick and Soll (2006) have documented that if a person wants to make a prediction and can also obtain the advice of an expert, he or she is often better off averaging both their opinions as opposed to differentially weighting one or the other. The underlying rationale for the power of averaging several judgments, forecasts, or variables, is simple. Basically, imagine that a prediction by one of *k* forecasters can be expressed as

$$z_j = \mu + \delta_j + \varepsilon_j \qquad (1)$$

where μ represents the overall average of all *k* forecasters; δ_j represents any bias specific to forecaster *j* ; and ε_j is an idiosyncratic error term associated with forecaster *j*. Now, if one simply assumes that δ_j and ε_j are uncorrelated and have means of zero across the *k* forecasters, it follows that taking the arithmetic average is an optimal

strategy (since the expected value of the criterion is equal to μ). Clearly such assumptions will not hold perfectly but, even if they are approximately true, the arithmetic average is a powerful predictor.

It is puzzling why people have such trouble in appreciating the power of the mean but perhaps this also explains, in part, why there is still such a belief in finding *the* different weights in regression analysis.

4. Discarding relevant information or when "less" can be "more." In normative theories of choice, the values of alternatives are typically assessed by calculating a weighted sum of outcomes. Thus, in expected utility theory, the utilities of outcomes are weighted by their probabilities of occurrence. Similarly, in the additive form of multi-attribute utility theory, the utility of an alternative $y_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is determined by the function.

$$U(y_i) = \sum_{j=1}^{k} w_j u(x_{ij}) \qquad (2)$$

where $U(.)$ denotes utility and the w_j are weighting parameters subject to the constraint that $\sum_{j=1}^{k} w_j = 1$ (see, e.g., Keeney & Raiffa, 1993).

Models such as (2) have a "gold standard" status in decision making because they essentially define what is "optimal." Moreover, they seem to make good sense in that they consider all the information and weight it appropriately. But do people need to consider all the information when they make a decision? Could they actually do "better" if they ignored some information?

One of the first researchers to examine this issue was Thorngate (1980). Using simulations, Thorngate investigated how often various heuristic strategies would select the highest expected value alternatives from different choices sets. In short, the criterion was a weighted sum (i.e., similar to equation 2 above) and the heuristic models

only used part of this information. For example, the most successful strategy in the simulation was one that assumed all probabilities were equal. Thorngate's results were surprising in that the more successful models had success rates of 75% and more when selecting the best from two to four alternatives. Clearly, for models to be effective, it was not necessary to use all the information.

Payne, Bettman, and Johnson (1993) conducted more simulations of the same type but also specifically considered the extent to which different heuristics involved varying levels of effort (conceptualized by the number of mental operations used in implementing them). These investigators also used the criterion of a weighted sum (e.g., similar to equation 2) and further investigated how different heuristics were subject to different task factors (e.g., levels of intercorrelations between variables and the relative presence/absence of dominated alternatives in choice sets). Once again, several heuristics that did not use all available information performed quite well. However, as in Thorngate's study, no heuristic could possibly perform better than the weighted sum of all information that was used as the criterion.

The conclusion from these studies was that heuristics could perform quite effectively but could never be better than using all information (indeed, this was how the studies were constructed). However, would it be possible to remove this design constraint and observe situations where "less" was "more"? Moreover, whereas one might justify models that use less information by accepting an accuracy-effort tradeoff, are there situations where one does not have to make this tradeoff?

In a 1996 paper, Gerd Gigerenzer and Daniel Goldstein indicated two ways in which "less" might be "more." Significantly, both involve the use of a heuristic decision rule exploiting an environmental "niche" (or task) to which it is well adapted.

The first example involves the use of the *recognition* heuristic (see also Goldstein & Gigerenzer, 2002).

At the extreme, imagine two people who have to choose between two alternatives. One person knows very little about the situation but does recognize one of the alternatives. She therefore chooses it. The second person, on the other hand, recognizes both alternatives and is generally quite knowledgeable about them. Normally, one would expect the second person to be more likely to make the correct choice. However, imagine that the first person's recognition knowledge is fairly highly correlated with the criterion. As the second person cannot use recognition to discriminate between the alternatives, he must use his additional knowledge. But, if he cannot use this additional knowledge to discriminate between the alternatives more accurately than the first person's "recognition knowledge," his choice will be less accurate. Paradoxically, although the first person has "less" knowledge, her predictive ability is "more" than that of the first.

The second phenomenon illustrated by Gigerenzer and Goldstein was the surprising predictive ability of the "take the best" heuristic (TTB). This is a simple, lexicographic decision rule for binary choices where selection depends on the first piece of information examined that discriminates between alternatives (information or cues are, however, consulted in the order of their predictive ability). When choosing between options characterized by binary attributes or cues, TTB is remarkably predictive and typically uses only a fraction of the information available. In the tests conducted by Gigerenzer and Goldstein (see also Gigerenzer, Todd, and the ABC Research Group, 1999), TTB generally outperforms equal weighting (that uses all the information, see above) and even regression models on cross-validation.

Gigerenzer and Goldstein did not claim that the predictive ability of TTB would hold in all environments and thus their demonstration was more by way of a "possibility." Nonetheless, although the effectiveness of TTB-like models have been demonstrated in areas of medical decision making (Breiman, Friedman, Olshen, & Stone, 1993), it is not clear that the implications have been realized to the advantage of both patients and physicians (i.e., faster and more accurate diagnoses). In medicine, in particular, professionals would appear to want to be seen to examine all information even if unnecessary.

Research, however, has not stopped at simply noting that less can be more. Importantly, different researchers have explored the conditions under which TTB and generalizations thereof are effective. Martignon and Hoffrage (1999; 2002) showed that, when the environment weights cues in a non-compensatory manner, TTB has optimal properties (by non-compensatory is meant that the implicit importance weight attached to each variable is greater than the sum of the weights of all variables less important than it). Natalia Karelaia and I showed that, for problems involving between three and five cues (or attributes), and choices involving between two and five alternatives TTB, and its generalization DEBA, are also effective over wide ranges of compensatory functions (Hogarth & Karelaia, 2005a; in press, a. See also Baucells, Carrasco, & Hogarth, 2006). Moreover, theoretical analyses done with a variety of simple models (where cues or attributes are both binary and continuous) show the general effectiveness of TTB-like models as well as illustrating further "less is more" effects (Hogarth & Karelaia, 2005b; in press, b).

In short, there are environments in which simple models can perform well relative to so-called "optimal" benchmarks. The key to understanding when this occurs lies in matching the features of simple rules with the demands of the environments in

which they operate (cf., the discussion of Makridakis's results above). From our work, we now know that much depends on: first, how "nature" or the environment weights cues; second, the amount of data available to estimate the "true" model; third, the level of redundancy amongst the cues; fourth, the amount of error in the environment; and fifth, the assumed loss function or how errors are penalized. Fortunately, all these factors can be quantified and it is possible to develop analytical results showing when particular simple rules do or do not work well (Hogarth & Karelaia, 2006; in press, b).

Being "rational" therefore involves having sufficient knowledge to know what to do in particular circumstances (i.e., matching one's decision rule to the demands of the environment). Noting, however, that often people may not know precisely what to do, Karelaia (2006) has suggested the use of strategies that hedge against one's lack of knowledge. Using both simulation and theoretical analyses, she has shown that one such strategy (that she calls CONF) performs quite well relative to other rules such as TTB or equal-weighting across several task environments (Hogarth & Karelaia, 2006; Karelaia, 2006).

It is interesting to note that many of these results contradict the intuitions of experts. For example, I made a presentation on this topic in a poster session at a professional conference attended by many leading researchers in decision analysis (the BDRM conference held at Duke University in 2004). Instead of presenting results, I created a competition by asking people to guess results given descriptions of simple environments and decision rules. There was also a $20 prize for the best set of estimates. Guesses, even by experienced decision analysts, did not match reality. There was considerable underestimation of the effectiveness of the simple models.

Empirically, "less is more" effects have been demonstrated and, theoretically, reasons why and when this occurs have been established for some cases. Perhaps it is

too soon to say yet how the scientific community will react to these findings. Based on past experience, the best bet is that it will take some time before these ideas are accepted.

Concluding comments

As the evidence reviewed above indicates, people – both in science and everyday life – are slow to accept evidence that challenges their beliefs and particularly when they have a stake in the latter. At one level, I see this as the inevitable consequence of a dilemma that has to be managed continuously by all living systems. This is the simultaneous need to adapt to change and yet maintain continuity and stability across time. Moreover, adapting to perceived change can involve two kinds of errors (i.e., adapting when one should not, and not adapting when one should) and the costs of error are not necessarily symmetric. Thus, without trying to rationalize what might seem to be dysfunctional behavior, it is legitimate to ask what conditions favor the adoption of new ideas that challenge the status quo and what, if anything, scientists can do to improve present practice.

From a descriptive viewpoint, economic incentives play an important role. For example, from the forecasting case study above, it is clear that practitioners in industry accept the implications of the time-series competitions even though theoretical statisticians might not share their enthusiasm. For scientists and others not faced by direct economic incentives, preserving reputation seems to be the greatest concern. The paradox, however, is that scientists who acknowledge that their theories are mistaken should – in principle – enhance their long-term reputations as scientists. Instead, there seems to be a larger short-term concern to preserve the status quo.

Two related suggestions have been made to overcome these difficulties. Some twenty years ago, Hofstee (1984) suggested that scientists engage in a system of reputational bets. That is, scientists with contradictory theories can jointly define how different outcomes of a future experiment should be interpreted (i.e., which theory is supported by the evidence). In Hofstee's scheme, the scientists assess probabilistic distributions over the outcomes (thereby indicating "how much" of their reputational capital they are prepared to bet) and a third, independent scientist runs the experiment. The outcomes of the experiment then impact on the scientists' reputational capitals or "ratings." However, I know of no cases where this system has actually been implemented.

A similar scheme involves a proposal labeled "adversarial collaboration." Here again, the disagreeing parties agree on what experiments should be run. An independent third party then runs the experiment which all three publish jointly. Unfortunately, it is not clear that this procedure resolves disputes. The protagonists may still disagree about the results (see, e.g., Mellers, Hertwig, & Kahneman, 2001).

Possibly one way to think about the situation is to use the analogy of the market place for ideas where, in the presence of efficiency, ideas that are currently "best" are adopted quickly. However, like real markets in economics and finance, the market for scientific ideas is not necessarily efficient. There are many situations where the market is "thin" and not all traders (i.e., scientists) have access to information. There are speculative "bubbles" or fashions as some theories become extremely popular for a time and then fade away (consider what happened to many learning models in psychology or applications of chaos theory in the social sciences). Great rewards are also to be had for identifying certain ideas that could become popular (e.g., cold fusion) and this too could distort information that is made public. Finally, despite attempts made to regulate the

exchange of ideas and the rules for doing science, people still find ways to circumvent regulations. In the final analysis, the market for scientific ideas can only become efficient in a long run sense. Unfortunately, as implied in a famous statement by Lord Keynes, our lives do not extend that far.

Finally, it is important not to consider the previous paragraph as suggesting a pessimistic cynicism. Each generation does see scientific progress and the accessibility of information has increased exponentially in recent years. The road to enlightenment, however, is bumpy.

References

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics*, 3, 63-84.

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, 134, 596-602.

Baucells, M., Carrasco, J. A., & Hogarth, R. M. (2006). *Cumulative dominance and heuristic performance in binary multi-attribute choice.* Working paper, IESE, UPC, & UPF, Barcelona.

Bazerman, M. (1997). *Judgment in managerial decision making* (4th ed.). New York, NY: Wiley.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis, forecasting, and control*. San Francisco, CA: Holden-Day.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees*. New York, NY: Chapman & Hall.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. *Educational and Psychological Measurement*, 32, 311-322.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist,* 34, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668-1674.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130 (6), 959-988.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance,* 7, 86-106.

Einhorn, H, J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13, 171-192.

Fair, R. C. (1986). Evaluating the predictive ability of models. In Z Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (pp. 1979-1995). Amsterdam: North-Holland.

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. Minneapolis, MN: University of Minnesota Press.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review*, 65, 289-308.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences,* 21, 199-200.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 650-669.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence of improving on clinical inferences. *Psychological Bulletin*, 73(6), 422-432.

Goldberger, A. S. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75-90.

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, 3, 398-400.

Greene, W. H. (1991). *Econometric analysis*. New York, NY: Macmillan.

Griffiths, W. E., Hill, R. C., & Judge, G. G. (1993). *Learning and practicing econometrics*. New York: Wiley.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 19-30.

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica*, 56, 93-109.

Hogarth, R. M. (1972). *Process tracing in clinical judgment: An analytical approach*. Unpublished PhD dissertation, The University of Chicago.

Hogarth, R. M. (1974). Process tracing in clinical judgment. *Behavioral Science*, 19, 298-313.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21, 40-46.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL: University of Chicago Press.

Hogarth, R. M., & Karelaia, N. (2005a). Simple models for multi-attribute choice with many alternatives: When it does and does not pay to face tradeoffs with binary attributes. *Management Science*, 51(12), 1860-1872.

Hogarth, R., M., & Karelaia, N. (2005b). Ignoring information in binary choice with continuous variables: When is less "more"? *Journal of Mathematical Psychology*, 49, 115-124.

Hogarth, R., M., & Karelaia, N. (2006). *On heuristic and linear models of judgment: Mapping the demand for knowledge.* Working paper, UPF, Barcelona.

Hogarth, R. M., & Karelaia, N. (in press, a). "Take-the-best" and other simple strategies: Why and when they work "well" with binary cues. *Theory and Decision*.

Hogarth, R. M., & Karelaia, N. (in press, b). Regions of rationality: Maps for bounded agents. *Decision Analysis*.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology*, 56, 1-12.

Holt, R. R. (1962). Individuality and generalization in the psychology of personality: A theoretical rationale for personality assessment and research. *Journal of Personality*, 30, 405-422.

Holt, R. R. (2004). A few dissents from a magnificent piece of work. *Applied & Preventive Psychology*, 11, 43-44.

Johnston, J. (1991). *Econometric methods* (3rd edition). New York, NY: McGraw-Hill.

Karelaia, N. (2006). Thirst for confirmation in multi-attribute choice: Does search for consistency impair decision performance? *Organizational Behavior and Human Decision Processes*, 100 (1), 128-143.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value tradeoffs*. Cambridge, UK: Cambridge University Press.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas. *Psychological Bulletin*, 107, 296-310.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, 56, 28-55.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions. Chicago, IL: The University of Chicago Press.

Larrick, R., & Soll, J. (2006). Intuitions about combining opinions: Misappreciation of the averaging rule. *Management Science*, 52 (1), 111-127.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently observed evidence. *Journal of Personality and Social Psychology,* 37, 2098-2109.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: An empirical investigation (with discussion). *Journal of the Royal Statistical Society. Series A*, 142, 97-145.

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, 451-476.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9, 5-23.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G Gigerenzer, P. M. Todd & the ABC Research Group. *Simple heuristics that make us smart* (pp. 119-140). New York: Oxford University Press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52, 29-71.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? *Psychological Science*, 12 (4), 269-275.

Mittelhammer, R. C., Judge, G. G., & Miller, D. J. (2000). *Econometric foundations*. New York, NY: Cambridge University Press.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of the Royal Statistical Society. Series A*, 137, 131-165.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.)*. New York, NY: McGraw-Hill.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY: Cambridge University Press.

Priestley, M. B. (1979). Discussion of the paper by Professor Makridakis and Dr. Hibon. *Journal of the Royal Statistical Society. Series A*, 142, 127-128.

Sawyer, J. (1966). Measurement *and* prediction: Clinical *and* statistical. *Psychological Bulletin*, 66, 178-200.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 31, 699-714.

Showers, J. L., & Chakrin, L. M. (1981). Reducing uncollectible revenues from residential telephone customers. *Interfaces*, 11, 21-31.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, 25, 219-225.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83, 213-217.

Wilks, S. S. (1938). Weighting schemes for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3, 23-40.

Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological Science*, 4, 58-62.