

Elsevier Editorial(tm) for Journal of Mathematical Psychology
Manuscript Draft

Manuscript Number:

Title: Descriptive Models as Generators of Prior Beliefs with Known
Equivalent Number of Observations (ENO)

Article Type: Regular Article

Keywords: Minimum variance rule, GRE subject exam, Effect size, Stevens' law, Prospect
theory, Hot hand.

Corresponding Author: Dr. Ido Erev Columbia University

Other Authors: Alvin E Roth, PhD; Robert L Slonim, PhD; Greg Barron, PhD Harvard
University, Case Western Reserve University, Harvard University

Dear Editor:

Attached please find the paper “Descriptive Models as Generators of Prior Beliefs with Known Equivalent Number of Observations (ENO)” that Al Roth, Bob Slonim, Greg Barron and I wish to publish in the Journal of Mathematical Psychology.

Sincerely,

Ido Erev

**Descriptive Models as Generators of Prior Beliefs with Known
Equivalent Number of Observations (ENO)**

Ido Erev

Minerva Center for Cognitive Studies and Faculty of Industrial Engineering and Management, Technion,

Alvin E. Roth

Harvard Business School and Department of Economics, Harvard University

Robert L. Slonim

Department of Economics, Case Western Reserve University

Greg Barron

Harvard Business School

Aug 10, 2004

Please address all correspondence to Ido Erev (erev@tx.technion.ac.il). We are very grateful for helpful conversations with David Budescu, Gary Chamberlain, Paul Feigin, Dave Krantz, Jack Porter, Tom Wallsten, Wolfgang Viechtbauer, and Richard Zeckhauser. We thank Maya Feldgoren for help in programming and running the experiments.

Abstract

When descriptive models are used to predict behavior in a novel condition, the predictions can be treated as “objective prior beliefs.” The current paper shows that an extension of the generalization criterion methodology proposed by Busemeyer and Wang (2000) can be used to facilitate objective updating of these beliefs. The optimal weighting of the initial prediction and the new data can be summarized with a single number: the model’s equivalent number of observations (ENO). The value of this measure is demonstrated with the analysis of models in three domains: Decision making, Direct perception, and Basketball.

Key words: Minimum variance rule, GRE subject exam, Effect size, Stevens’ law, Prospect theory, Hot hand.

1. INTRODUCTION

It is often useful to treat the predictions of quantitative models of human behavior as “prior beliefs.” For an example consider the prediction of the performance of a particular student in college. The best psychometric model is expected to provide a valid initial prediction that can be used as the college’s prior beliefs. When the performance of the student in the first semester is observed, this belief can be updated in an attempt to improve the prediction of the student’s future performance.

Another example involves the prediction of choice behavior in a decision problem that has not previously been studied. The decision making model that provides the best predictions in similar problems (that have been studied) can be used to generate a good initial prediction. And this prediction can be improved as data from a direct examination of the relevant problem becomes available.

Initial predictions generated by good quantitative models have two desirable properties. First, when the models are appropriately selected, these predictions are likely to be relatively accurate. Second, the process of generating these predictions is clear and objective; different users of the model are expected to generate similar predictions. The main goal of the current paper is to propose and evaluate a method that facilitates similar accuracy and objectivity during the updating of the initial predictions.

The basic idea behind the proposed method can be described as an extension of Busemeyer and Wang’s (2000) generalization criterion methodology for model selection. The method involves a two-step estimation that prescribes the initial theory-based prediction, and the best weighting of this initial prediction with the accumulating observations. The initial prediction can be derived by an estimation of the model parameters for the relevant population of possible situations. The optimal weighting of the model’s prediction with the mean of the first direct observations can be estimated using a set of regression equations.

We show that the optimal weighting of the initial prediction and the new data can be summarized with a single number: the model’s *equivalent number of observations* (ENO). This number is closely related to commonly used statistics like Student’s t statistic and Cohen’s measure of effect size. The relationship of the ENO statistic to known measures clarifies the

similarities and differences between basic research that focuses on model comparison, and the application of these models as “objective prior beliefs.”

To demonstrate the proposed procedure, this paper estimates the equivalent number of observations of models in three domains: decision-making, direct perception, and basketball. The examples show that the ENO of simple models can be rather large. In addition our results show that some of the best-studied factors seem to have little effect on the ENO of the models we examined, while other factors that have received little attention have large effects.

2. THE PROPOSED PROCEDURE

We consider a situation in which a specific model (a point prediction rule) provides clear predictions to a well-defined population of experimental conditions. Experimental condition i is associated with a true mean μ_i and some distribution around this mean. The model’s (rule’s) prediction of the mean of condition i is $R_i = \mu_i + \alpha_i$ where α_i is an error term (symmetrically distributed around 0).

During the “prediction task,” one of the experimental conditions is randomly selected. The user of the model can observe m independent draws from the condition’s distribution and is asked to predict the next draw.

Notice that the information available to the user can be summarized with two numbers: the initial model’s prediction, and the mean of the first m observations. Both numbers are examples of noisy estimates (of the population mean). Thus, given the necessary data, finding their optimal (least squared distance, or according to other criteria) weighting is reduced to a familiar problem with a well-known solution in forecasting statistics.

Assume that the user of the model is allowed to analyze a data set that includes the results of 100 conditions, each with $n = 50$ observations, drawn from the same population of experimental conditions. To use the available statistical methods it is convenient to organize the data set in a matrix with 5000 (50X100) rows. Each row involves a particular observation in a particular condition (a dependent variable) and two predictors. The dependent variable x_{ij} describes (is) observation j in condition i . The predictors are the two estimates: R_i is the

prediction of the model for condition i, and \bar{X}_{oj} is the mean of other 49 observations (all observations but j) in condition i. Table 1 presents a smaller data set of this type.

<Insert Table 1>

The optimal (least squared difference) weighting of the two estimates can be calculated with a regression analysis:

$$(1) \quad x_{ij} = \beta_0 + \beta_1(R_i) + \beta_2(\bar{X}_{oj}) + \varepsilon_{ij}$$

2.1 Implications for experimental design

The attempt to extend the example presented above to actual prediction tasks reveals two difficulties. First, most theoretical models (including expected utility theory) do not make quantitative point predictions. Second, even if we could derive the quantitative predictions of our favorite model, we require an appropriate data set to compute the optimal weights.

We make a distinction between theories and theory-based point prediction rules. We use the term “point prediction rules” to refer to rules that make ex ante quantitative predictions for a well-specified set of conditions. Under this definition the derivation of point prediction rules from most theories requires the addition of auxiliary assumptions, and parameter estimation. For example, to derive point predictions from expected utility theory (see Section 3.1) one has to assume a particular functional form and then estimate the parameters (or distribution of parameters).

In some cases it is possible to use specific point prediction rules proposed in previous research. For instance, Tversky and Kahneman’s (1992) work suggests a point prediction rule based on cumulative prospect theory, certain simplifying assumptions, and parameters estimated using a certainty equivalent method based on the median behavior in their cash equivalent study. Notice that there are many reasons why point prediction rules are likely to be inaccurate: The theory might be incorrect and/or oversimplified, the auxiliary assumptions might not be descriptive, and the estimation might not be accurate. For the current analysis the source of the inaccuracy is not important.

The first step in the experimental procedure we propose calls for the estimation of parameters that will allow the theory in question to be transformed into a point prediction rule for a well-defined set of situations. To make clear that we are not estimating parameters to fit some historical set of experimental observations, and in order to estimate parameters that will be useful in predicting behavior on unknown examples from the universe in question, we estimate parameters on behavior observed on a random sample of tasks from the universe of conditions from which those to be predicted later will be drawn.

The second step of the experimental procedure involves the estimation of the optimal weighting of the point prediction and new data. Because we are interested in how the point predictions should be weighted with multiple experimental observations of an as yet to be determined task, the estimation of combination weights is done on a second random sample of tasks, with n independent observations taken for each task. Just like the data described in Table 1, this second data set can be used to run the regression analysis.

A shortcoming of the two-samples experimental design is its cost. If our goal is to predict the behavior in a particular condition, running two experiments on different conditions seems cumbersome. Nevertheless, if a particular theory is expected to have many useful predictions and it is not known in advance what the specific conditions are that will be addressed, the cost may not be too large.

2.2 Relationship to Bayesian statistics

To see the relationship of the suggested regression-based procedure to Bayesian statistics recall that in the current setting the minimum variance rule, used to address similar problem in Bayesian statistics, yields the same outcome as regression analysis with the constraint $\beta_0 = 0$, $\beta_1 + \beta_2 = 1$ (see Granger and Ramanathan, 1984; Gupta and Wilton, 1987).¹ Under the minimum variance rule (and restricted regression analysis) the optimal weight for the point prediction rule is

$$(2) \quad \hat{W} = \hat{\beta}_1 = \frac{MSE(\bar{X}_o) - CD(\bar{X}_o, R)}{MSE(\bar{X}_o) + MSE(R) - 2CD(\bar{X}_o, R)}$$

where $MSE(R)$ is an unbiased estimator of the mean squared error (MSE) of the point prediction rule, $MSE(\bar{X}_o)$ is an unbiased estimator of the MSE of the second predictor (the mean of the other observations), and $CD(\bar{X}_o, R) = r(\bar{X}_o - x_{ij}, R - x_{ij})\sqrt{MSE(\bar{X}_o)MSE(R)}$ is the common deviation ($r(\bar{X}_o - x_{ij}, R - x_{ij})$ is the correlation between the deviations). To clarify the meaning of these terms, Table 1 shows how they are computed in this 30-observations example.

As demonstrated by Granger and Ramanathan (1984) this constraint impairs the accuracy of the combined estimate (relative to unconstrained regression). Yet, in a small-sample setting, the simpler one-parameter combination rule can be useful as it reduces the risk of over-fitting the data.

2.3 The Equivalent Number of Observations (ENO)

Under the minimum variance the two estimates receive equal weight when $MSE(\bar{X}_o) = MSE(R)$. The number of observations used to derive the experiment-based predictions ($m = n-1$), decreases the error of this prediction ($MSE(\bar{X}_o)$), but does not have a systematic effect on the error of the model-based prediction ($MSE(R)$). Thus, it is possible to estimate the size of the experiment for which the two predictors are equally useful and the optimal weight is 0.5. We refer to this value as the model's estimated Equivalent Number of Observations (ENO). Appendix 1b shows that the exact value² is given by,

$$(3) \quad ENO = S^2 / (M - S^2).$$

where S^2 is the pooled variance (over tasks in the experiment), and $M = MSE(R)$.

Note that the ENO is a property not of the model alone but of both the model and the data. On a universe of tasks over which subjects exhibit little variance in behavior, every

¹ Under the Bayesian interpretation the model is a summary of our prior belief.

² When the samples are small it is possible to obtain estimates such that $M - S^2 \leq 0$. We interpret this inequality to imply that the data suggest that the ENO of the model is too high to be estimated based on the current sample. In restricted regression analysis, these cases lead to the estimation of $\beta_1 > 1$. In Erev, Roth, Slonim & Barron (2002), the quantity called predictive value, derived in a different way, is equal to ENO.

observation is very informative, so even a very good model will have a low ENO. This point is clarified in Section 2.5.

When the errors of the two predictors are not correlated (that is, $CD(\bar{X}_o, R) = S^2$), the ENO statistic can be used to provide a simplified approximation of the optimal weighting.³ The implications of the minimum variance rule to the current setting (after observing m subjects) are reduced (see Appendix 1b) to the assertion:

$$(4) \quad \hat{W} = \hat{\beta}_1 = \frac{ENO}{ENO + m}$$

Equation 4 makes clear why we refer to the “equivalent number of observations” of a theoretical prediction, since, if $ENO = k$, we give the theoretical prediction the same weight as we would a data set of k observations, when combining it with a data set with m observations.

2.4 ENO in a particular task and the relationship to Students t and Cohen’s d .

Although the main goal of the procedure proposed here is to analyze populations of conditions, it is easy to see that the ENO statistic can be written as a transformation of conventional statistics used in analysis of individual conditions. Most importantly, the ENO statistic based on one condition ($N=1$) and a sample of n observations is closely tied to the t -test for the hypothesis that the condition’s mean equals 0 (the model $R = 0$). As shown in Appendix 1c:

$$(5) \quad t^2 = \frac{n}{ENO} + 1$$

Equivalent number of observations is similarly related to Cohen’s (1994) *effect size* (d). Since the null hypothesis is not likely to be exactly correct, an increase in the sample size increases the

³ To see why $CD(\bar{X}_o, R) = S^2$ implies independent error note that $CD(\bar{X}_o, R)$ can be written as the sum of the mean squared deviation of x_{ij} from the relevant population’s mean (this error equals S^2), and the common error of the two predictors in predicting the population’s mean. Notice that in Table 1’s small data set example the estimated value of $CD(\bar{X}_o, R)$ is larger than S^2 (112.5 versus 98.9). In the data sets we considered the deviation between the estimates decreases when more observations are used in the estimation.

t value and eventually leads to rejecting the null hypothesis. The sensitivity of the t statistic to sample size implies that it does not say much about the effect size (the extent to which the model

is wrong). Cohen proposed d as an alternative to the t statistic: $d = \frac{|R - \bar{X}|}{S}$ where R is the

model's (rule) point prediction, \bar{X} is the mean of the observations, and S is the standard deviation of the observations. Since $d^2 = (t^2)/n$ the relation of d to ENO is given by

$$(6) \quad d^2 = \frac{1}{ENO} + 1/n$$

This equation allows d to be interpreted in terms of the number of observations needed to outperform the null hypothesis (that the point prediction rule in question is correct). For example, Cohen argues that $d=1$ is a “large” effect. We can now see that $d = 1$ implies an ENO (of the null hypothesis) of only a little more than one observation, when n is large.

Similarly the connection to the t -test makes clear that any model with a finite ENO can be rejected given a large enough sample. There is an approximately constant ratio between the model's ENO and the sample size needed for rejection. Given a “significance level” of .05 (t^2 around 4) the ratio is 1 to 3. For instance, a model (null hypothesis) with ENO of 100 will be rejected given a sample size of 300 observations at the .05 level ($t_{399} = 2$) but not at the .01 level. Thus, computation of ENO can be used to provide a new interpretation of hypothesis testing statistics (see Krantz, 1999 for a similar idea).

Finally, when n is large ENO converges to the ratio of the population variance and the rule's squared bias $(R-\mu)^2$ where μ is the population mean.

2.5 ENO and the Proportion of variance accounted for by the model

In order to clarify the relationship between ENO and traditional measures of predictive validity, like the proportion of variance accounted for by a model, it is convenient to consider psychometric predictions. Psychometric models, typically built on regression analysis (see

review in Anastasi, 1996), provide valuable point predictions of the expected success of individuals in particular tasks (e.g., performance in College or in a new work environment).

Assume that the grade of student i in exam j is given by

$$(7) \quad G_{ij} = \mu_i + \varepsilon_{ij} = R_i - \alpha_i + \varepsilon_{ij}$$

where μ_i is the true mean ε_{ij} is the error, R_i is the prediction of the rule (psychometric model) for the performance of student i , and α_i is the model's bias in the predictions of student i .

When the different terms are drawn from independent normal distributions with known variances, the proportion of variance accounted for by the psychometric model (in predicting the first exam) is:

$$(8) \quad \rho^2 = \frac{\sigma_R^2}{\sigma_R^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2}$$

where σ_k^2 is the variance of the k th term.

Consider now the task of predicting performance of a particular student in one course based on the pre admission psychometric model, and the performance in a different course. The expected ENO of the psychometric model (in predicting Exam n of a particular individual based on the first $n-1$ exams of this individual) is:

$$(9) \quad ENO = \frac{\sigma_\varepsilon^2}{M - \sigma_\varepsilon^2} = \frac{\sigma_\varepsilon^2}{(\sigma_\alpha^2 + \sigma_\varepsilon^2) - (\sigma_\varepsilon^2)} = \frac{\sigma_\varepsilon^2}{\sigma_\alpha^2}$$

Thus, ENO and the proportion of accounted variance (validity) are only partially related: Both measures decrease with the systematic bias σ_α^2 , but the error parameter σ_ε^2 has different effect in the two cases. Moreover, ENO ignores the effect of σ_R^2 .

To clarify this limited relationship, consider the following two numerical examples. Example “Large random error” is defined by the parameters: $\sigma_R = 1$, $\sigma_\alpha = 0.5$, $\sigma_\varepsilon = 2$. Example “Large student bias” is defined by the parameters: $\sigma_R = 1$, $\sigma_\alpha = 2$, $\sigma_\varepsilon = 0.5$. It is easy to see that the model accounts for the same proportion of the variance in the two cases ($\rho^2 = 0.19$). Yet, the model has very different “within student ENO” in the two environments. When the random error variability is large, the ENO is 16: the psychometric model is as accurate as the mean of 16 exams taken by the relevant student. The suggested weighting of the model and the first exam in predicting the second exam is $(16/17)(R_i) + (1/17)(\text{first exam})$. However, when the student bias is small, the model’s ENO drops to 1/16. In this environment the suggested weighting of the model and the first exam in predicting the second exam is $(1/17)(R_i) + (16/17)(\text{first exam})$.

An extreme example involves the case of an almost perfect model with zero error variance. For instance: $\sigma_R = 1$, $\sigma_\alpha = 0.01$, $\sigma_\varepsilon = 0$. The model in this example accounts for almost all the variance ($\rho^2 = 0.9999$). Nevertheless, its ENO is 0, because even one observation will give a perfect prediction since there is zero variance.

These examples illustrate that ENO is not an alternative to ρ^2 and similar measures of predictive validity. These measures capture an important property of models -- their value in predicting the first observation (and when updating of the initial prediction is not allowed). ENO is not a very useful measure of this property. ENO’s main value is in complementing ρ^2 -like measures by capturing the weight that should be given to the initial prediction when it is combined with additional information.

2.6 Relationship to model selection.

As noted above the procedure proposed here is an extension of the generalization criterion proposed by Busemeyer and Wang (2000) as a model selection methodology (see review in Forster, 2000 and Myung, 2000). The extension involves the computation of the optimal weighting of the estimated model with new data. This extension complicates the analysis and does not add to the value of the generalization criterion as a model selection tool. In other words, the current procedure is not an efficient model selection tool. It is designed to facilitate the updating of initial beliefs generated by the model that has been selected.

3. EMPIRICAL EXAMPLES

This section presents three examples of applying the ENO statistic. The first example considers a basic research in which the two-stage analysis is needed. It illustrates the estimation of point prediction rules of popular models, and their ENOs. The other two examples focus on existing point prediction rules.

3.1 Example 1: Decision making under uncertainty

Decision making under uncertainty is probably the best-studied problem in behavioral decision research and experimental economics. Following von Neumann and Morgenstern's (1947) axiomatization of expected utility, most studies focus on binary choice among simple gambles. This line of research has discovered robust violations of choices predicted by expected utility theory, and led to the development of elegant descriptive alternatives. Nevertheless, very little is known about the value of these models as generators of prior beliefs. In an attempt to address this issue, the current analysis estimates the ENO of leading models on a space of simple choices. It focuses on situations in which the choices faced by subjects are all of the following form: choose one of a pair of alternative gambles, each gamble having two possible outcomes, one zero and the second positive.

Seventy-six subjects (38 from Harvard and 38 from the Technion) chose among one hundred pairs of gambles from one of two random samples. Let (v_g, p_g) be a gamble that pays v_g with probability p_g and 0 otherwise. The pairs of gambles (conditions) studied here were created by random sampling of the four relevant values v_1, p_1, v_2, p_2 . The probabilities were randomly sampled from the uniform distributions of $[0.00, 0.01, 0.02, \dots, 1.00]$ and the values were sampled from the set $[1, 2, 3, \dots, 100]$. To eliminate trivial choices the final two samples were limited to pairs meeting the constraint $(v_1 - v_2)(p_1 - p_2) < 0$ (i.e. the gamble with the bigger prize has the smaller probability of winning). Table A1 in Appendix 2 shows the two sets of gambles. In the experiment the pairs of gambles were presented on the computer screen, one after another, with subjects 'clicking' on their preferred gamble from each pair. Subjects were paid a show-up fee and received the outcome of one of their chosen gambles, selected randomly. The exchange

rate was 1 point = 2.5 cents. No feedback was provided during the choice stage. The payoff determining gambles were selected and played after the subjects completed all their choices.

3.1.1 Expected value and nine models:

We focus on five basic deterministic models: the expected value (EV) rule, a power expected utility rule (PEU), and three versions of cumulative prospect theory (CPT, Tversky & Kahneman, 1992; Tversky & Wakker, 1995; Prelec, 1998; Gonzalez & Wu, 1999).⁴ In addition, we explore stochastic variants of these models. All of these models evaluate a gamble (v_g, p_g) by a function of the form $\pi(p_g)U(v_g)$. The models differ with respect to the assumed functions and response rule.⁵

The basic models assume a deterministic selection of the alternative with higher weighted value. The EV rule uses the identity functions $\pi(p_g) = p_g$ and $U(v_g) = v_g$. PEU uses the identity probability function and a power utility function $U(v_g) = (v_g)^\alpha$ where $0 < \alpha < 1$. In the current context (gain domain) the original version of CPT (CPT1) assumes $U(v_g) = (v_g)^\alpha$ and $\pi(p_g) = p_g^\gamma / (p_g^\gamma + (1 - p_g)^\gamma)^{1/\gamma}$. Following Prelec (1998) the second variant of CPT (CPT1P) assumes an exponential weighting function: $\pi(p_g) = \text{EXP}(-\ln(p)^\gamma)$. The third variant of CPT, referred to as CPT2 (following Gonzalez & Wu, 1999 and Lattimore et al., 1992) has a 2-parameter weighting function: $\pi(p_g) = \delta p_g^\gamma / (\delta p_g^\gamma + (1 - p_g)^\gamma)$.

Following Busemeyer (1985) the stochastic models assume the same functions as the basic model, but replace the deterministic response rule with the following probabilistic rule:

$$(10) \quad P(I) = \frac{e^{q_1 \lambda / s}}{\sum_{g=1}^2 e^{q_g \lambda / s}}$$

⁴ In the current setting (one non zero outcome) the predictions of the configural weight models like TAX (Birbaum & Chavez, 1997) are indistinguishable than the prediction of PT.

⁵ In the current settings there is no reason to distinguish between theories based on the motivation behind their basic assumptions. "Normative" theories can be used to derive descriptive point prediction rules. See Rapoport and Wallsten (1972) for a discussion of the difference between the motivation for the basic assumptions of a model (experimental or normative) and the precision of its predictions.

where q_g is the weight of lottery g , λ is a strength of preference parameter and s (the normalization factor) is the average distance between the cumulative utility functions of the two lotteries:

$$(11) \quad s = \text{Abs}[U(v_1) - U(v_2)]\text{Min}(\pi(p_1), \pi(p_2)) + \text{Min}(U(v_1), U(v_2))\text{Abs}[\pi(p_1) - \pi(p_2)]$$

To clarify the computation of s Figure 1 presents the cumulative utility function of the two gambles in Problem 1 (60, 0.80; 0) or (74, 0.75; 0). Since $U(74) > U(60)$ and $\pi(0.80) > \pi(0.74)$, the absolute distance in this example is

$$(12) \quad s = [U(74) - U(60)]\pi(0.75) + U(60)[\pi(0.80) - \pi(0.75)]$$

<Insert Figure 1>

3.1.2 The predictions and the observed behavior

The current data sets can be analyzed in many ways. In particular, the two samples of subjects (Harvard and Technion) can be pooled or analyzed individually, and each one of the two samples of lotteries can be used to estimate parameters or to compute ENO. These different ways lead to almost identical outcomes. Thus, the current section focuses on the most natural analysis: pooling over the two universities, estimating parameters with the first sample of gambles (Form 1) and estimating regression weights and ENOs for the models on the second sample (Form 2). The main experimental results (proportion of Left choices) are summarized in Table A1.

The left-hand side of Table 2 presents a summary of the mean squared deviation scores (M) of the ten point prediction rules and the pooled estimated variance (S^2). The ENO column (center) shows the equivalent number of observations of the different predictions in the second sample. The results show a large advantage for the stochastic models. Even a 1-parameter stochastic model (stochastic EV) outperforms all the deterministic models. In addition, the results show that the assumption of a stochastic response rule is particularly effective given prospect theory valuation rules. Whereas CPT does not outperform PEU, the stochastic variants

of CPT show a large advantage over the stochastic version of PEU. In one case the stochastic response rule multiplies the ENO of prospect theory by more than nine.⁶

<Insert Table 2>

The three types of combinations of the point predictions with the new (Form 2) data, discussed above, are compared in Table 2⁷: unconstrained regression (3 parameters), constrained regression (1 parameter), and ENO based weighting. This analysis was conducted for three levels of $n-1$ (the number of other subjects used to predict each subject): 1, 3 and 37. The results support the following conclusions: (1) Constrained regression and ENO weighting lead to practically identical predictions and fit. The differences are smaller than 0.1% of the M score in all cases. Given this similarity the results of the two analyses were collapsed to one column in Table 2. This similarity suggests that the errors of the two estimates are approximately independent. Namely $CD(\bar{X}_o, R) \approx S^2$. (2) As shown in the right side of Table 3 adding parameters to the regression analysis improves the fit. (3) The advantage of the 3-parameter combination decreases with n . For $n-1 = 37$ the improvement achieved with the 3-parameter regression is significant; the constraint $\beta_0 = 0, \beta_1 + \beta_2 = 1$ cannot be rejected.

To further evaluate the difference between the ENO weighting and unconstrained regression, we estimated the weighting parameters of the different combination rules based on the first 50 gambles (of Form 2) and compared the predictions of the last 50 gambles. The results support the conclusions from hypothesis testing. For the SCPT models the ENO combination slightly outperforms the combination implied by unconstrained regression in the current data set. This observation suggests that in this case the constrained weighting reduces the risk of over-

⁶ In additional analyses we obtained the following results: (1) The power version of EU outperforms exponential and logarithmic versions. (2) With the two-parameter version of Prelec's (1998) weighting the ENO of SCPT is 21.5. (3) With the median parameters of CPT estimated by Tversky & Kahneman (1992) the model has ENO below 1.

⁷ In addition to the analyses summarized in the right-hand column of Table 2 we also examine combination based on logistic regression. The logistic regression predictions are associated with slightly higher mean squared error scores than the scores of the linear regression, and do not change the main results.

fitting the data.⁸ Obviously, however, we can be confident that given very large samples (more than 100 conditions) unconstrained regression will reliably improve the predictions.

In summary, then, the current analysis provides clear recommendations on how to apply each of the predictive models we consider to the Harvard square problem. For example, the optimal weighting of the SCPT2 prediction with the behavior of the first m subjects in predicting the behavior of subject $m+1$ is obtained with the equation $P(\text{Left}) = w\text{Pr}(\text{SPT}) + (1-w)P(\text{other } m)$ with $w = 23.13/(23.13+m)$, where $\text{Pr}(\text{SPT})$ equals the prediction of SCPT2 and $P(\text{other } m)$ is the observed proportion of Left choices by the first n subjects.

In addition the current analysis demonstrates that CPT, which was developed to summarize observations from selected experimental problems (like the Allais (1953) paradox) chosen to illustrate systematic failures in expected utility theory, does not provide a good point prediction for randomly selected problems. Yet, a relatively minor addition to this model, the assumption of a rank dependant stochastic response rule, substantially improves its equivalent number of observations for new problems.

3.2 Example 2: Stevens' Power Law

In a series of elegant studies Stevens and his co-workers (see review in Stevens, 1975) have demonstrated that a simple model, often referred to as Stevens' power law, can be used to provide useful predictions of subjective estimates of objective magnitudes. According to Stevens' power law the estimate Ψ of the objective magnitude ϕ is:

$$(13) \quad \Psi = k\phi^\beta$$

where k is a normalization factor, and β is a sensory continuum parameter. Thus,

$$(14) \quad \log(\Psi) = \beta \log(\phi) + \log(k)$$

⁸ One interesting source for over fitting involves the naming of the gambles as Left or Right. The estimates of unconstrained regression are sensitive to this manipulation (the estimated intercept is effected by the attractiveness of the gambles named Left). The constrained regression is robust to this manipulation.

The research that supports Stevens power law has been criticized for overlooking demonstrations of significant deviations from the basic law. For example, many studies have documented range effects that are ignored by the basic model (see e.g., Poulton, 1968). Moreover, within-subject analysis seems to suggest that the power function is an artifact of grouping (Pradhan & Hoffman, 1963). Stevens addressed this criticism with the assertion that the deviations reflect measuring noise, and modeling them might mask the robust principle captured by the power law.

The interpretation of models as objective prior beliefs can be used to clarify and evaluate Stevens' assertion. Specifically, the logic behind Stevens' assertion can be understood as the suggestion that the Power law has very high ENO. Thus, the fact that it is not exactly correct does not imply that it should be rejected. In order to evaluate this suggestion we run a simple experiment that compute Stevens law's ENO in the context of visual area estimation.

Fourteen Technion students participated in the experiment. The participants were told that their task is to estimate the area of squares that will be presented on the computer screen. They were presented with a "standard" stimulus of a 10x10 centimeter square and were told that the area of this standard is "100." The twenty stimuli presented in Table 3 were judged in the experiment. Implicit in the current selection of tasks is the assumption that when the space of tasks is narrow (involves only one dimension with clear bounds) the representative sample can be more efficient than a random sample. The order of the different stimuli was independently randomized for each participant.

<Insert Table 3>

Stevens' (1975) estimate of the parameter β in visual area estimation is 0.7. The current instructions imply a normalization factor $k = 100/(100^{0.7}) = 3.98$. Thus, the point predictions of Stevens' law are given by:

$$(15) \quad \log(\Psi) = (0.7)\log(\phi) + \log(3.98).$$

Table 3 shows the predicted and observed log judgments. The logarithmic transformation is used to avoid the power law artifact (see Anderson & Tweney, 1997; and Myung, Cheongtag, & Pitt, 2003). The results show that this point prediction rule is indeed very useful. Its ENO is 40.9!

3.3 Example 3: Predicting free shots

An interesting variant of the sequential psychometric prediction task considered in Section 2.5 involves the prediction of the probability that a certain NBA player will hit his next free throw attempt. Previous research (Gilovich et al., 1985) suggests that in the NBA the long-term (career) hit rate of a player is a good predictor (model) of the hit probability on his next shot. Yet the hot hand literature does not imply that players do not have good and bad days (see Dorsey-Palmateer & Smith, 2004). It is reasonable to assume that the hit rate of a player during a particular game may add to our ability to predict the next shot. This assumption is consistent with the conventions developed in televised broadcasting of games. Both statistics are typically presented on the screen when a player is on the line trying to hit a free throw.

The ENO measure provides an efficient summary of the relevant information for this estimation problem. In this example it is reasonable to think about a class of games played by a particular player as the relevant space of tasks. For example, all the regular season games played by Shaquille O'Neal is a class that should be taken into account by the coaches of the opposing teams. The coaches' goal is to derive the optimal weighting of the player's career hit rate with the observed hit rate in the current game. The current analysis suggests that this goal can be easily performed by calculating the players' career average ENO and using the weight $W = \text{ENO}/(\text{ENO} + k)$ during the game (where k is the number of observed free throw attempts in the current game).

To demonstrate that this analysis leads to interesting results Table 4 presents the relevant ENO scores of the six players who led the NBA in free throw attempts during the 2000/2001 season. Games with two or less attempts were omitted from the analysis. The results are based on the simplifying assumptions of no sequence within a game and equal weight to each game. That is, all the shots in a particular games were treated as symmetrical observations.

<Insert Table 4>

Table 4 shows that the career average has high (9-44) ENO. Yet, using the first k shots in a game is likely to improve the predictions of future shots. The median ENO over the six players is around 19.

Notice that there are two potential reasons for the limited ENO of the long term average: Variation from season to season, and variations from game to game. Analysis of Tim Duncan's data highlights this difference. He had a particular bad free throw season (average of 0.616 relative to career average of 0.71). In addition, he had some very bad days (including a 0 hit from 7 tries) and very good days. To evaluate the contribution of the within season variation the right most column in Table 4 presents the ENO of a model that uses the season average as predictor. The results show a moderate increase in ENO for four players, and large increase for the other two. In one case (Allan Iverson) the ENO of the season average is too high to be estimated from the available data.

4. INTERPRETATION, AND SOME CAUTIONS

Perhaps the most striking thing about the preceding analysis is how low are the ENO's of predictors like expected value, expected utility, and deterministic prospect theory. These numbers need to be interpreted with caution, keeping in mind the observation (derived in Section 2.5) that ENO complements and does not replace measures of predictive validity. This observation has two important implications.

The first implication is that unlike typical measures of predictive validity, ENO increases with the variance of the data. As the definition, $ENO = S^2 / (M - S^2)$, makes clear, if the variance in the data is low, no prediction that is not precisely correct can have a high ENO. But, for given variance, the closer the prediction is to the data (i.e. the smaller is M), the higher the ENO.

The second implication is that the ENO determines the weight that should be attached to the prediction *after* we have at least one observation from the task to be predicted. A prediction with a low ENO might nevertheless be a reasonable predictor *before* any data are gathered. This point is of course related to the preceding point: if the variance in the data is low, even one

observation of the specific task to be predicted may be quite a valuable predictor, and ENO is a measure of what the theoretical prediction adds to the predictive power of even a small set of highly relevant observations.

That being said, consider the simplest of the predictions we have examined in Section 3.1, Expected Value (EV), on the set of gambles we used to compare it to other theory-based point prediction rules (Table A1, sample 2). The first thing to note is that 77% of the choices made by subjects are consistent with expected value. That is, more than three quarters of all the choices observed are choices of the gamble with the higher expected payoff. (And 83% of all choices were consistent with each of the other deterministic point prediction rules.) Yet on those random pairs of gambles, the ENO of Expected Value is barely over 1 (and that of the other deterministic rules is less than 2.5, see Table 2).

To understand what is going on, note that, of the 100 pairs of gambles in sample 2, there are 44 such that at least 90% of the subjects made the same choice. And of these 44 pairs of gambles with very low variance, only 1 has almost all subjects choosing the gamble with the *lower* expected value.⁹

At the same time it is important to recall that Section 3.1's choice data does not reflect the extreme condition analyzed in the last paragraph of Section 2.5 (zero variance that implies high predictive validity and ENO=0). On the actual data set we considered, of course, the variance is positive. (On only 6 of the 100 gamble pairs in sample 2 did 100% of the subjects make the same choice.) So, the deterministic theory-based point prediction rules based on utility theory and prospect theory have higher ENO's than does expected value because they do better on the hard cases. And the considerable advantage of the stochastic models is easy to understand, because they predict more closely the *proportion* of subjects who will make the choice predicted by the related deterministic rule, and not merely which choice will be made by most subjects. The fact that the stochastic rules in Table 2 have ENO's in the double digits indicates that, even when a dozen or more observations of the task to be predicted have already

⁹ In gamble pair #23 in sample 2, the gambles are $g1 = (36, .35)$ with expected value 12.6, and $g2 = (13, .95)$ with expected value 12.35, and only 2 out of 38 subjects chose gamble 1, i.e. 95% chose gamble 2. In contrast, the other deterministic prediction rules predict the second gamble will be chosen, e.g. the PEU of these two gambles, with $\alpha = .39$, is $PEU(g1) = 1.42$ and $PEU(g2) = 2.58$, while the three versions of cumulative prospect theory as parameterized in Table 2 evaluate $g1$ and $g2$ respectively as: 1.00 and 1.47 (CPT1); 1.40 and 3.03 (CPT2); 0.97 and 1.82 (CPT1P).

been observed, there is enough variance so that a theoretical point prediction rule can still improve the prediction made about future observations.

So in interpreting ENO's, it is important to remember that they are not a property of a theory's predictions in isolation, but are a function of the theory and the data set. In low variance data, each observation is very informative, and even predictions that are close to the data may have low ENO's.¹⁰ However on a given data set, predictions that are closer to the data have higher ENO's.

5. SUMMARY

The current paper highlights the value of quantitative models as generators of prior beliefs. Beliefs generated by good quantitative models have two desired properties: they are objective and relatively accurate. The main contribution of the present paper is the development of a simple procedure that allows objective and optimal (under certain assumptions) updating of these model-based initial predictions. The proposed procedure requires elaborate two-stage estimation analysis, but can be summarized with a simple measure with clear relationship to known statistics. In addition, this summary has an intuitive interpretation: it is the model's equivalent number of observations (ENO).

The basic arguments presented here are not limited to models of human behavior. Nevertheless, there are good reasons to believe that the ENO statistic is likely to be most useful in facilitating the use of behavioral models. One reason involves the coexistence (in the behavioral sciences) of imprecise models, relatively cheap observations that can be used to update initial predictions, and large situation effect on the value of the initial observations (the magnitude of S^2). These properties imply that potential users of behavioral models are likely to obtain observations that deviate from the model's predictions, but cannot know if the deviation reflects a biased model (that can be ignored once the first observations are available) or large observation variability. A prior computation of the ENO of the model can reduce this difficulty.

¹⁰ For example, if you are in a bar in a country whose language you can't read, you may nevertheless have a good guess about which of the restrooms is for men and which is for women. However you have only to see one or two native men enter one of the doors before your prior prediction becomes unimportant.

A second and related reason involves the suggestion that behavioral models are underused. An interesting support for this suggestion comes from analysis of the GRE subject exams in Psychology and in Physics (these exams are used by leading universities to evaluate candidates for graduate school, see <http://www.gre.org/edindex.html>). Most questions in physics ask the examinee to predict behavior in well-defined situations. A typical question might show a simple circuit and ask “if the switch is closed at time zero, which of the following curves shows the current through the resistor as a function of time?” In psychology, on the other hand, almost all questions ask the examinee to state the meaning of a particular term, or associate it with a particular investigator.¹¹ We believe that this difference is driven by the fact that the developers of the exams and the universities that use them do not believe that a good understanding of leading psychological theories help answer questions that require predictions. We hope that the computation of the ENO of leading models will help change this common belief and reduce the gap between Psychology and Physics.

¹¹ See <http://www.gre.org/subjtest.html> for sample tests in both physics and psychology. Erev and Livne-Tarandach (2004) classified these questions to three categories: Experiment-based, abstract and mixed. Their analysis reveals that the percentage of abstract questions is 9% in Physics, and 84% in Psychology.

REFERENCES

- 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.
- Anastasi, A. (1996). *Psychological testing* (7th ed.). New York: Macmillan.
- Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior & Human Decision Processes*, 71, 161-194.
- Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed sample, and sequential sampling models." *Journal of Experimental Psychology*, 11, 538-564.
- Busemeyer, J. R., & Wang, Y. M., (2000). Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology*, 44 (1), 171-189.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997-1003
- Dorsey-Palmateer, R., & Smith, G. (2004). Bowlers' Hot Hands
<http://www.economics.pomona.edu/GarySmith/bowling/bowling.html>
- Erev, I., & Livne-Tarandach, R. (2004). Experiment-based exams and the difference between the behavioral and the natural sciences. Working paper.
- Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88, 4, 848-881.
- Erev, I., Roth, A. E., Slonim, S. L., & Barron G. (2002). Predictive value and the usefulness of game theoretic models. *International Journal of Forecasting*. 18(3), 359-368.
- Forster, M. R. (2000). Key Concepts in Model Selection: Performance and Generalizability. *Journal of Mathematical Psychology*. 44(1), 205-231.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17, 295-314.
- Gonzalez R., & Wu G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38, 129-166.
- Granger, C. W. & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3, 197-204.
- Gupta, S., & Wilton P. C. (1987). Combination of forecasts: An extension. *Management Science*, 33(3), 356-372.

- Haruvy, E., & Erev I. (2001). On the applications and interpretation of parameters in learning models.”
In Zwick R. & Rapoport A. *Experimental Business Research*. (Kluwer Academic Publisher), pp
285-300.
- Kahneman, D., & Tversky A. (1979). Prospect theory: An analysis of decisions under certainty.
Econometrica, 47, 263-291.
- Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American
Statistical Association*, 94 (448): 1372-1381.
- Lattimore P.K., Baker J. R., Witte A. D. (1992). The influence of probability on risky choice—a
parametric examination. *Journal of Economic Behavior and Organization*, 17 (3): 377-400.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical
Psychology*. 44(1), 190-204.
- Myung, I. J., Cheongtag, K., & Pitt, M. A. (2000). Toward an explanation of the power law artifact:
Insights from response surface analysis. *Memory and Cognition*. 28(5), 832-840.
- Pradhan, P. L., & Hoffman, P. J. (1963). Effect of spacing and range of stimuli on magnitude estimation
judgments. *Journal of Experimental Psychology*, 66, 533-541.
- Poulton, E. C. (1968). The new Psychophysics: Six models for magnitude estimation. *Psychological
Bulletin*, 17, 139-147.
- Prelec, D. (1998). The probability weighting function. *Econometrica*, 66, 497-527.
- Rapoport, A., & Wallsten T. S. (1972). Individual decision behavior. *Annual Review of Psychology*, 23,
131-176.
- Stevens S. S. (1975). *Psychophysics*. New York, NY: Wiley.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of
uncertainty. *Journal of Risk and Uncertainty*, 9, 195-230.
- Tversky, A., & Wakker, P. (1995). Risk attitude and decision weights. *Econometrica*, 63, 1255-1280.
- von Neumann, J., Morgenstern, O. (1947). *The Theory of Games and Economic Behavior*, Princeton, NJ:
Princeton Univ. Press.

Table 1: Example of a data set that allows evaluation of the optimal combination of a model, and a data-based estimate. Variable x_{ij} summarizes the value of observation j in condition i , \bar{X}_{oij} is the mean of other (two) observations, R_i is the prediction of the model. The right hand columns and the lower rows present the computation of the statistics used in the current analysis.

Condition	Obs.	Depend. variable	Predictors		Deviations		Squared deviations	
			x_{ij}	\bar{X}_{oij}	R_i	$(\bar{X}_{oij} - x_{ij})$	$(R_i - x_{ij})$	$(\bar{X}_{oij} - x_{ij})^2$
1	1	100	116	104	16	4	256	16
	2	110	111	104	1	-6	1	36
	3	122	105	104	-17	-18	289	324
2	1	124	109	92	-15	-32	225	1024
	2	106	118	92	12	-14	144	196
	3	112	115	92	3	-20	9	400
3	1	102	114	108	12	6	144	36
	2	110	110	108	0	-2	0	4
	3	118	106	108	-12	-10	144	100
4	1	126	117	124	-9	-2	81	4
	2	106	127	124	21	18	441	324
	3	128	116	124	-12	-4	144	16
5	1	62	77	76	15	14	225	196
	2	74	71	76	-3	2	9	4
	3	80	68	76	-12	-4	144	16
6	1	130	136	128	6	-2	36	4
	2	134	134	128	0	-6	0	36
	3	138	132	128	-6	-10	36	100
7	1	104	100	90	-4	-14	16	196
	2	112	96	90	-16	-22	256	484
	3	88	108	90	20	2	400	4
8	1	70	94	79	24	9	576	81
	2	90	84	79	-6	-11	36	121
	3	98	80	79	-18	-19	324	361
9	1	92	99	94	7	2	49	4
	2	90	100	94	10	4	100	16
	3	108	91	94	-17	-14	289	196
10	1	84	86	76	2	-8	4	64
	2	82	87	76	5	-6	25	36
	3	90	83	76	-7	-14	49	196
					$r(\bar{X}_o - x_{ij}, R - x_{ij}) = 0.746$		Sum = 4452	Sum = 4595
$MSE(\bar{X}_o) = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{X}_{oij})^2 = \frac{1}{10} \frac{1}{3} 4452 = 148.4$								
$MSE(R) = M = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n (x_{ij} - R_i)^2 = \frac{1}{10} \frac{1}{3} 4595 = 153.17$								
$S^2 = MSE(\bar{X}_o) \left(\frac{n-1}{n} \right) = 148.4 \left(\frac{2}{3} \right) = 98.93$, $ENO = S^2 / (M - S^2) = 98.93 / (153.17 - 98.93) = 1.82$								
$CD(\bar{X}_o, R) = 0.746 \sqrt{(148.4)(153.17)} = 112.5$								

Table 2: The M (mean squared error) scores and ENO of different point predictions in Example 1. Sample 1 was used for parameter estimation and Sample 2 was used for estimating regression weights and ENO . The right hand columns compare different combinations of the model and the data (S before the model's name implies a stochastic response rule).

Point prediction rule (underlying theory and parameters)	M sample 1 (fitted)	M sample 2	ENO ($S^2 = 0.12$)	Unconstrained regression (3 parameters)				Constrained regression (1 parameter) and ENO based weighting*		
				n-1	1	3	37	1	3	37
EV (expected value)	0.229	0.230	1.09	β_0	.162	.096	.0125	.515	.267	.030
				β_1	.357	.219	.030			
				β_2	.326	.592	.946			
				M	.159	.142	.123			
PEU (power expected utility) $\alpha=0.39$	(0.173)	0.169	2.44	β_0	.142	.111	.022	.708	.451	.063
				β_1	.564	.439	.090			
				β_2	.148	.338	.865			
				M	.138	.133	.123			
CPT1 (Cumulative prospect theory, original version) $\alpha=0.33, \gamma=0.75$	(0.167)	0.169	2.44	β_0	.142	.109	.022	.708	.451	.063
				β_1	.564	.441	.090			
				β_2	.148	.338	.865			
				M	.138	.133	.123			
CPT1p (CPT with $\pi(p_g) = \text{EXP}(-\ln(p)^\gamma)$) $\alpha=0.28, \gamma=0.72$	(0.167)	0.171	2.34	β_0	.142	.109	.021	.708	.451	.061
				β_1	.564	.441	.086			
				β_2	.148	.338	.872			
				M	.138	.133	.123			
CPT2 (CPT with $\pi(p_g) =$ $\delta p_g^\gamma / (\delta p_g^\gamma + (1 - p_g)^\gamma)$. $\alpha=0.4, \gamma=0.8$ $\delta=0.7$	(0.167)	0.169	2.44	β_0	.142	.109	.022	.708	.451	.063
				β_1	.564	.439	.090			
				β_2	.148	.338	.865			
				M	.138	.133	.123			
SEV $\lambda=2$	(0.162)	0.159	3.08	β_0	-.038	-.026	-.005	.760	.503	.075
				β_1	.826	.546	.085			
				β_2	.237	.493	.924			
				M	.149	.140	.123			
SPEU $\alpha=0.36, \lambda=2.3$	(0.131)	0.129	12.91	β_0	-.035	-.030	-.011	.932	.815	.258
				β_1	1.012	.891	.297			
				β_2	.064	.173	.727			
				M	.128	.128	.122			
SCPT1 $\alpha=0.34, \gamma=0.72$ $\lambda=2.7$	(0.127)	0.126	20.68	β_0	-.023	-.020	-.009	.956	.876	.358
				β_1	.989	.906	.379			
				β_2	.042	.121	.634			
				M	.126	.125	.122			
SCPT1p $\alpha=0.33, \gamma=0.68$ $\lambda=2.7$	(0.127)	0.169	22.50	β_0	-.028	-.046	-.012	.959	.870	.378
				β_1	1.00	.960	.407			
				β_2	.038	.115	.609			
				M	.125	.124	.122			
SCPT2 $\alpha=0.5, \gamma=0.8$ $\delta=0.5, \lambda=2.6$	(0.127)	0.125	23.13	β_0	-.028	-.025	-.012	.961	.886	.384
				β_1	1.00	.928	.414			
				β_2	.037	.110	.603			
				M	.125	.124	.122			

* ENO weighting and 1 parameter constrained regression yield the same results to three significant figures, and so are presented together.

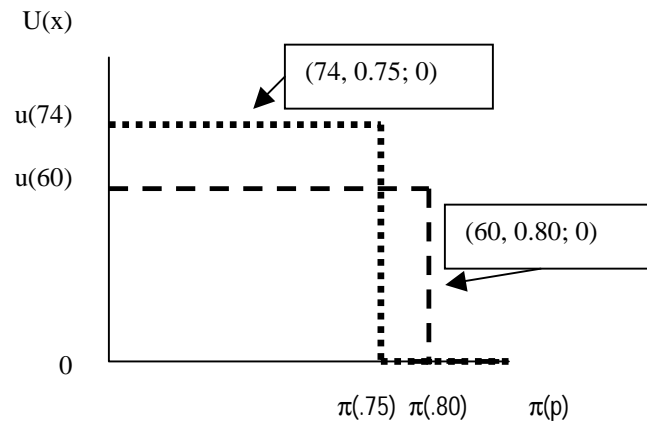
Table 3: The twenty squares studied in Section 3.2 are presented on the left column. The predictions (in natural logarithm) were derived based on the parameter $\beta = 0.7$ estimated by Stevens (1975), and the normalization factor implied by β and the current procedure. The estimated ENO is 40.9.

Square dimensions	Objective area	Prediction	Mean log estimate	S^2	M
1x1	1	1.38	1.51	0.75	0.77
2x2	4	2.35	2.37	0.69	0.69
3x3	9	2.92	3.14	0.19	0.24
4x4	16	3.32	3.42	0.19	0.20
5x5	25	3.63	3.64	0.19	0.19
6x6	36	3.89	3.91	0.16	0.16
7x7	49	4.11	4.08	0.14	0.15
8x8	64	4.29	4.37	0.05	0.05
9x9	81	4.46	4.44	0.06	0.06
10x10	100	4.61	4.56	0.02	0.03
11x11	121	4.74	4.67	0.02	0.03
12x12	144	4.86	4.84	0.07	0.07
13x13	169	4.97	4.91	0.16	0.17
14x14	196	5.08	5.06	0.20	0.20
15x15	225	5.17	5.16	0.12	0.12
16x16	256	5.26	5.23	0.12	0.12
17x17	289	5.35	5.29	0.14	0.14
18x18	324	5.43	5.41	0.16	0.16
19x19	361	5.50	5.50	0.21	0.21
20x20	400	5.58	5.57	0.25	0.25
pooled				0.1958	0.2006

Table 4: Predicting free throws by six NBA players based on Career average or Season average (the models) and new observation (k other free throw attempts in the game).

Player	Career average	Number of Games	Attempts	Season average	S^2	Career average		Season average
						M	ENO	ENO
Shaquille O'Neal	0.534	74	972	0.513	0.238	0.250	20.29	21.05
Jerry Stackhouse	0.788	75	804	0.823	0.139	0.147	17.94	21.39
Paul Pierce	0.772	76	727	0.748	0.185	0.189	44.08	44.50
Allan Iverson	0.723	68	714	0.815	0.151	0.159	19.77	V. Large
Karl Malone	0.735	74	666	0.794	0.160	0.167	24.76	54.31
Tim Duncan	0.710	74	647	0.617	0.221	0.245	9.034	13.41

Figure 1: The cumulative utility functions of the two gambles in Problem 1 in Table 1: $(60, 0.80; 0)$ or $(74, 0.75; 0)$. The normalization factor s is the average distance between the two functions.



APPENDIX 1:

a. The derivation of ENO

Equal weight is implied when $MSE(\bar{X}_o) = MSE(R)$

$$\widehat{MSE}(\bar{X}_o) = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{X}_{oij})^2$$

where x_{ij} is the choice of subject j in task i , and \bar{X}_{oij} is the mean of the “other $n-1$ subjects” in this tasks.

Notice that

$$\bar{X}_{oij} = \frac{n\bar{X}_i - x_{ij}}{n-1}$$

where \bar{X}_i is the mean of all n subjects in task i . These equations imply

$$MSE(\bar{X}_o) = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n \left(x_{ij} - \frac{n\bar{X}_i - x_{ij}}{n-1}\right)^2 = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n \frac{n^2}{(n-1)} \frac{(\bar{X}_i - x_{ij})^2}{(n-1)} = \frac{n}{(n-1)} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n \frac{(\bar{X}_i - x_{ij})^2}{(n-1)} = \frac{n}{(n-1)} S^2$$

where S^2 is the pooled error variance.

$MSE(R) = M$ is the estimate of the mean squared distance between the rule’s prediction and each observation:

$$M = \frac{1}{N} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n (x_{ij} - R_i)^2$$

Thus, equal weight for predicting the next observation based on $n-1$ observations is therefore expected when

$$\frac{n}{(n-1)} S^2 = M$$

and, the implied value of $n-1$ (the size of the experiment expected to yield prediction as accurate as the model) is $S^2/(M - S^2) = ENO$.

b. The optimal weighting as a function of ENO

When $CD(\bar{X}_o, R) = S^2$:

$$\hat{W} = \hat{\beta}_1 = \frac{MSE(\bar{X}_o) - S^2}{MSE(\bar{X}_o) + M - 2S^2}$$

Replacing $MSE(\bar{X}_o)$ with $\frac{n}{(n-1)} S^2$ we get:

$$\hat{W} = \hat{\beta}_1 = \frac{\frac{n}{(n-1)}S^2 - S^2}{\frac{n}{(n-1)}S^2 + M - 2S^2} = \frac{S^2/n - 1}{S^2/n + M - S^2} = \frac{S^2}{S^2 + (n-1)(M - S^2)}$$

$$\frac{1}{\hat{W}} = \frac{S^2 + (n-1)(M - S^2)}{S^2} = 1 + (n-1)\frac{M - S^2}{S^2} = 1 + \frac{(n-1)}{ENO} = \frac{(n-1) + ENO}{ENO}$$

Thus,

$$\hat{W} = \hat{\beta}_1 = \frac{ENO}{(n-1) + ENO}$$

c. The relationship of ENO to the t statistic.

The definition of the t statistic implies

$$t^2 = \frac{(\mu - \bar{X})^2}{S^2/n}$$

Under the null hypothesis $\mu = 0$ (the model $R = 0$):

$$t^2 = \frac{\bar{X}^2}{S^2/n}$$

The definition of S^2 implies

$$S^2 = \frac{\sum_{j=1}^n (\bar{X} - x_j)^2}{(n-1)} = \frac{\sum_{j=1}^n (\bar{X}^2 - 2\bar{X}x_j + x_j^2)}{(n-1)} = \frac{\sum_{j=1}^n (x_j^2) - n\bar{X}^2}{(n-1)}$$

When the model states $R = 0$, and we consider only one task ($N=1$) $nM = \sum_{j=1}^n (x_j^2)$ Thus,

$$S^2 = \frac{nM - n\bar{X}^2}{(n-1)}$$

$$\bar{X}^2 = M - \frac{n-1}{n}S^2 = M - S^2 + S^2/n$$

Replacing for \bar{X}^2 in the definition of t^2 we get:

$$t^2 = \frac{M - S^2 + S^2/n}{S^2/n} = \frac{M - S^2}{S^2/n} + 1 = \frac{n}{ENO} + 1$$

APPENDIX 2: The random gambles, and zero-sum games

Table A1: The 200 gamble pairs and the observed proportion of L choices studied in Example 1

Sample (Form) 1	Left (L)		Right (R)		P(L)
	V1	P1	V2	P2	
1	90	0.24	55	0.77	0.05
2	23	0.30	57	0.09	0.82
3	1	0.75	69	0.45	0.21
4	36	0.99	77	0.77	0.55
5	13	0.87	45	0.19	0.92
6	24	0.37	93	0.27	0.24
7	67	0.08	35	0.63	0.03
8	11	0.63	94	0.33	0.32
9	45	0.25	29	0.64	0.13
10	79	0.93	31	0.98	0.87
11	62	0.69	82	0.53	0.74
12	76	0.10	46	0.63	0.05
13	42	0.19	37	0.66	0.00
14	37	0.32	86	0.08	0.89
15	87	0.69	81	0.76	0.18
16	84	0.87	90	0.04	0.97
17	13	0.99	68	0.70	0.58
18	77	0.49	98	0.17	0.87
19	13	0.93	21	0.75	0.74
20	81	0.88	66	0.98	0.26
21	59	0.99	68	0.15	0.97
22	37	0.11	32	0.66	0.05
23	51	0.55	70	0.12	0.95
24	48	0.34	21	0.35	0.89
25	98	0.06	52	0.79	0.08
26	69	0.08	33	0.47	0.05
27	2	0.69	70	0.47	0.16
28	82	0.31	35	0.61	0.29
29	52	0.91	10	0.95	0.82
30	23	0.25	83	0.05	0.76
31	75	0.24	76	0.08	0.92
32	55	0.44	82	0.12	0.87
33	78	0.99	93	0.44	0.97
34	28	0.51	65	0.48	0.24
35	9	1.00	55	0.86	0.34
36	79	0.87	43	1.00	0.45
37	5	0.12	95	0.03	0.29
38	74	0.04	52	0.45	0.03
39	10	0.36	7	0.92	0.05
40	43	0.05	7	0.54	0.24
41	58	0.90	70	0.06	0.92
42	62	0.79	80	0.49	0.95
43	21	0.46	86	0.38	0.26
44	94	0.28	51	0.96	0.08
45	78	0.05	33	0.39	0.11
46	11	0.12	3	0.66	0.08
47	44	0.11	36	0.79	0.00
48	98	0.02	91	0.76	0.00
49	95	0.50	74	0.82	0.13
50	29	0.39	9	0.82	0.42

#	Left (L)		Right (R)		P(L)
	V1	P1	V2	P2	
51	94	0.01	84	0.57	0.03
52	49	0.58	38	0.81	0.13
53	18	0.74	69	0.30	0.74
54	61	0.51	58	0.87	0.05
55	48	0.03	24	0.77	0.03
56	53	0.97	58	0.51	0.92
57	51	0.93	81	0.59	0.84
58	81	0.20	52	0.89	0.05
59	55	0.96	88	0.19	0.92
60	74	0.19	62	0.52	0.03
61	100	0.32	84	0.94	0.11
62	36	0.68	41	0.56	0.87
63	68	0.24	67	0.73	0.03
64	5	0.82	97	0.54	0.16
65	32	0.99	100	0.79	0.42
66	32	0.27	7	0.84	0.37
67	36	0.97	82	0.63	0.76
68	23	1.00	70	0.83	0.37
69	100	0.03	56	0.76	0.05
70	85	0.46	83	0.67	0.05
71	83	0.16	5	0.91	0.42
72	88	0.74	19	0.86	0.79
73	46	0.67	18	0.68	0.89
74	57	0.72	71	0.14	0.95
75	37	0.25	34	0.76	0.05
76	98	0.03	82	0.10	0.29
77	56	0.12	53	0.28	0.05
78	77	0.21	6	0.49	0.68
79	11	0.65	71	0.27	0.61
80	34	0.32	44	0.23	0.76
81	80	0.22	16	0.76	0.37
82	52	0.72	85	0.30	0.92
83	1	0.27	83	0.24	0.16
84	30	0.97	42	0.31	1.00
85	67	0.25	61	0.37	0.16
86	86	0.16	39	0.66	0.03
87	75	0.29	18	0.77	0.29
88	93	0.02	83	0.15	0.08
89	65	0.53	43	0.85	0.13
90	72	0.34	34	0.70	0.39
91	90	0.27	51	0.80	0.34
92	21	0.26	25	0.21	0.63
93	8	0.78	61	0.68	0.11
94	43	0.81	61	0.25	0.97
95	92	0.14	45	0.28	0.32
96	49	0.38	38	0.99	0.03
97	38	0.54	11	1.00	0.26
98	30	0.56	31	0.46	0.89
99	36	0.71	95	0.62	0.32
100	74	0.28	39	0.29	0.84

Table A1 continued:

Sample (Form) 2	Left (L)		Right (R)		P(L)
	V1	P1	V2	P2	
1	60	0.80	74	0.75	0.39
2	89	0.43	88	0.80	0.16
3	38	0.73	60	0.70	0.21
4	98	0.33	88	0.89	0.00
5	52	0.61	60	0.32	0.97
6	96	0.15	20	0.59	0.32
7	48	0.63	90	0.34	0.82
8	93	0.38	57	0.77	0.13
9	57	0.94	70	0.23	0.97
10	88	0.19	68	0.37	0.16
11	42	0.24	1	0.47	0.92
12	49	0.16	12	0.74	0.05
13	93	0.20	33	0.57	0.16
14	65	0.93	75	0.78	0.92
15	86	0.09	22	0.76	0.08
16	41	0.51	65	0.35	0.84
17	80	0.81	92	0.58	0.97
18	31	0.94	77	0.55	0.84
19	81	0.22	9	0.64	0.61
20	20	0.81	36	0.01	0.97
21	83	0.97	95	0.25	0.95
22	97	0.73	83	0.89	0.24
23	36	0.35	13	0.95	0.05
24	66	0.96	100	0.51	0.82
25	74	0.74	85	0.67	0.71
26	6	0.60	33	0.49	0.21
27	64	0.25	40	0.48	0.05
28	2	0.52	93	0.44	0.16
29	98	0.55	61	0.57	0.89
30	96	0.90	90	0.94	0.39
31	59	0.90	79	0.08	0.89
32	14	0.76	58	0.53	0.26
33	22	0.59	65	0.45	0.24
34	80	0.40	87	0.07	0.92
35	37	0.65	45	0.55	0.76
36	31	0.35	30	0.69	0.05
37	93	0.14	71	0.83	0.03
38	69	0.45	33	0.66	0.50
39	88	0.17	50	0.35	0.16
40	27	0.84	54	0.64	0.50
41	2	0.08	1	0.35	0.00
42	91	0.60	43	0.65	0.84
43	8	0.76	38	0.02	0.92
44	56	0.05	11	0.81	0.03
45	86	0.06	82	0.91	0.00
46	56	0.16	49	0.56	0.03
47	63	0.81	67	0.29	0.95
48	76	0.93	100	0.48	0.92
49	39	0.05	14	0.16	0.24
50	34	0.24	16	0.99	0.00

#	Left (L)		Right (R)		P(L)
	V1	P1	V2	P2	
51	85	0.10	2	0.69	0.53
52	8	0.61	7	0.97	0.08
53	17	0.43	91	0.13	0.66
54	16	0.94	18	0.14	0.92
55	91	0.53	9	0.63	0.84
56	1	0.80	88	0.71	0.03
57	100	0.24	79	0.81	0.05
58	39	0.01	14	0.22	0.18
59	31	0.93	55	0.72	0.66
60	11	0.95	79	0.62	0.42
61	79	0.24	19	0.91	0.29
62	72	0.29	78	0.13	1.00
63	25	0.33	94	0.03	0.84
64	70	0.78	52	0.82	0.71
65	19	0.97	94	0.51	0.50
66	36	0.17	41	0.13	0.45
67	36	0.44	10	0.88	0.21
68	92	0.57	47	0.67	0.76
69	5	0.88	69	0.84	0.08
70	93	0.55	31	0.60	0.92
71	68	0.52	1	0.93	0.68
72	85	0.12	32	0.15	0.95
73	7	1.00	31	0.51	0.50
74	90	0.45	13	0.54	0.92
75	66	0.17	24	0.67	0.08
76	78	0.82	3	0.89	0.89
77	49	0.07	35	0.90	0.03
78	78	0.46	63	0.94	0.03
79	26	0.48	46	0.06	1.00
80	37	0.90	48	0.63	0.68
81	87	0.66	96	0.47	0.89
82	21	0.89	96	0.87	0.13
83	46	0.70	13	0.77	0.79
84	75	0.07	57	0.10	0.39
85	74	0.09	30	0.28	0.03
86	83	0.32	59	1.00	0.13
87	54	0.95	86	0.17	0.97
88	79	0.23	21	0.79	0.13
89	16	0.63	70	0.54	0.08
90	57	0.74	60	0.39	0.97
91	62	0.07	46	0.53	0.05
92	30	0.93	40	0.37	0.97
93	98	0.57	75	0.79	0.26
94	39	0.78	78	0.70	0.24
95	57	0.79	82	0.12	0.95
96	75	0.19	4	0.45	0.76
97	18	0.40	71	0.30	0.18
98	41	0.18	91	0.10	0.29
99	36	0.55	99	0.06	0.87
100	33	0.97	87	0.17	0.92