Ignoring information in binary choice with continuous variables:

When is less "more"?

Robin M. Hogarth & Natalia Karelaia[*]

ICREA & Universitat Pompeu Fabra, Barcelona

robin.hogarth@upf.edu

natalia.karelaia@upf.edu

October 2004

Abstract

When can a single variable be more accurate in binary choice than multiple sources of information? We derive analytically the probability that a single variable (SV) will correctly predict one of two choices when both criterion and predictor are continuous variables. We further provide analogous derivations for multiple regression (MR) and equal weighting (EW) and specify the conditions under which the models differ in expected predictive ability. Key factors include variability in cue validities, intercorrelation between predictors, and the ratio of predictors to observations in MR. Theory and simulations are used to illustrate the differential effects of these factors. Results directly address why and when "one-reason" decision making can be more effective than analyses that use more information. We thus provide analytical backing to intriguing empirical results that, to date, have lacked theoretical justification. There are predictable conditions for which one should expect "less to be more."

Ignoring information in binary choice with continuous variables:

When is less "more"?

Many choices in life are binary in nature. You decide, for example, to hire one of two candidates; you choose one of two vacation sites or automobiles or movies; and so on. Normative theories recommend evaluating alternatives by multi-attribute models that specify both the attributes of choice as well as their relative importance (cf., Keeney & Raiffa, 1993). Thus, for example, the value or utility of an alternative $y_i = (x_{i1}, x_{i2}, ..., x_{ik})$ is determined by the function.

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

where $U(.)$ denotes utility and the w_j are weighting parameters subject to the constraint that $\sum_{j=1}^{k} w_j = 1$.

If experienced in a particular domain, a person's choice process might be accurately portrayed by models like equation (1). However, in many situations, people lack the experience necessary to perform analogous cognitive operations and resort to simpler mechanisms. This paper explores implications of such simpler mechanisms.

Our starting point is a remarkable set of studies by Gigerenzer and Goldstein (1996; Gigerenzer, Todd, & the ABC Group, 1999) in which a simple lexicographic model named "take-the-best" (TTB) proved to be highly predictive in binary choice. Across 20 datasets, cross-validated predictive accuracies were 71% for TTB, 69% for "equal weighting," and 68% for multiple regression.

TTB works as follows. First, it is assumed that attributes or cues can be ordered by their ability to predict the criterion. Second, choice is made by the most predictive cue that can discriminate between options. Finally, if no cues discriminate,

choice is made at random. This model is "fast and frugal" in that it typically decides on the basis of a few or even a single cue (Gigerenzer et al., 1999).

A possible limitation of these findings is that the attributes or cues were binary in nature (i.e., could only take values of 0 or 1). Perhaps, it could be argued, if variables had been continuous, TTB would not have been so successful in relative terms. Gigerenzer et al. (1999) tested this hypothesis on their datasets. In making TTB operational with what they called "exact quantitative values," they adopted a strict interpretation of discrimination. A cue discriminates if its value on one alternative exceeds that of the other by any amount. In effect, this means that, with continuous variables, TTB is based on the single most important variable. With "exact quantitative values" as opposed to binary variables, multiple regression did in fact improve performance – to 76%. But so did TTB which also achieved 76%.[1]

These findings raise the important issues of why and when single-variable models can predict as well as or better than models that use more information. In other words, when is less "more"? In related work, we have analyzed this question theoretically for binary cues and identified several factors including types of weighting functions (e.g., the distribution of weights, w_j, in equation 1) and specific characteristics of distributions of alternatives (Hogarth & Karelaia, 2004). See also Martignon and Hoffrage (1999; 2002).

Recently, Fasolo, McClelland, and Todd (2004) conducted an intriguing set of simulations of multi-attribute choice using continuous variables (involving 21 options characterized by 6 attributes). Their goal was to assess how well choices by models with differing numbers of attributes could predict total utility and, in doing so, they

---

[1] Gigerenzer et al. (1999) do not report the equivalent statistics for equal weighting. In addition, some of their datasets included categorical variables (e.g., presence or absence of a Bundesliga football team in a German city) that could not have taken continuous values.

varied levels of average intercorrelations among the attributes (positive, zero, and negative) and types of weighting functions (equal and differential). Their results showed important effects for both average intercorrelations and weighting functions. With differential weighting, one attribute was sufficient to capture at least 90% of total utility. With positive intercorrelation among attributes, there was not much difference between the results of equal and differential weighting. When intercorrelation was negative, however, equal weighting was sensitive to the number of attributes used (the more, the better).

The goal of this paper is to provide a theoretical analysis – for the case of continuous variables – of why and when single-variable models can be more effective in binary choice than models that use more variables (including the same single variable). The paper is structured as follows.

We first derive analytical expressions for the probability that a single variable (SV) will correctly discriminate between two specific alternatives on a given criterion. We show this to be an increasing function of the variable's predictive validity as well as the difference between the values of the variable exhibited by the alternatives. We further derive an analytical expression for the overall predictive ability of SV for populations of observations. Analogous expressions are then derived for equal weighting (EW) and multiple regression (MR). Comparisons of the expressions for SV, EW, and MR lead to expectations as to when particular models are likely to be more or less effective. The key factors are: (1) variance in cue validities; (2) intercorrelation between predictor variables; and (3) numbers of observations, n, and variables, k, used in estimating coefficients for MR. We illustrate the effects of these factors by both theoretical calculations and simulations. In brief, we find that SV is surprisingly effective relative to models that use more variables and this is especially

the case when predictor variables are positively intercorrelated.   Finally, we discuss

implications of our findings.

How accurate is a single variable for binary choice?

To assess expected accuracy of a single variable (SV) in binary choice,

consider choosing between alternatives from a distribution characterized by two

correlated random variables, one of which is a criterion, Y, and the other an attribute,

X. Furthermore, assume that alternative A is preferred over alternative B if $y_a > y_b$.[2]

Now, imagine that the only information about A and B are the values that they exhibit

on X.   Denote these specific values by x_a and x_b, respectively. Without loss of

generality, assume that $x_a > x_b$.   Whether A or B should be chosen can be

characterized by the probability that $Y_a > Y_b$, given that $x_a > x_b$. In other words, what

is $P\{Y_a > Y_b \mid X_{a=} x_a > X_{b=} x_b\}$?

To answer this question, assume that Y and X are both standardized normal

variables, i.e., both are $N(0,1)$.[3]   Moreover, the two variables are positively correlated

(if they are negatively correlated, simply multiply one by -1). Denote the correlation

by the parameter ρ_{yx}, ($\rho_{yx} > 0$). Given these facts, it is possible to represent Y_a and Y_b

by the equations:

$$Y_a = \rho_{yx} X_a + \varepsilon_a \tag{2}$$

$$\text{and} \quad Y_b = \rho_{yx} X_b + \varepsilon_b \tag{3}$$

where ε_a and ε_b are normally distributed error terms, each with mean of 0 and

variance of $(1 - \rho^2_{yx})$, independent of each other and of X_a and X_b.

---

[2] In what follows, we denote random variables by upper case letters, e.g., Y and X, and specific values
or realizations by lower case letters, e.g., y and x. As an exception to this practice, we use lower case
Greek letters to denote random error variables, e.g., ε.
[3] We discuss the normality assumption in the discussion section below.

Using equations (2) and (3), the difference between Y_a and Y_b can be written as

$$Y_a - Y_b = \rho_{yx}(X_a - X_b) + (\varepsilon_a - \varepsilon_b) \tag{4}$$

Thus, $Y_a > Y_b$ if

$$\rho_{yx}(X_a - X_b) > \varepsilon_b - \varepsilon_a \tag{5}$$

The question of determining $P\{Y_a > Y_b \mid X_{a=}x_a > X_{b=}x_b\}$ can now be reframed as determining the probability that the right hand side of (5) is smaller than $\rho_{yx}(X_a - X_b)$. As can be seen, this latter term is the product of ρ_{yx}, the correlation between Y and X, and $(X_a - X_b)$ or the difference between X_a and X_b. In other words, the larger the correlation between Y and X, and the larger the difference between X_a and X_b, the greater $P\{Y_a > Y_b \mid X_{a=}x_a > X_{b=}x_b\}$.

To determine this probability, we make use of the fact that the difference $(\varepsilon_b - \varepsilon_a)$ is normally distributed with mean of 0 and variance of $2(1 - \rho_{yx}^2)$. The properties of the normal distribution can therefore be used to calculate

$$P\{Y_a > Y_b \mid X_a = x_a > X_b = x_b\} = P\{(\varepsilon_b - \varepsilon_a) < \rho_{yx}(x_a - x_b)\} =$$

$$\int_{-\infty}^{s^*} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \tag{6}$$

where $s^* = \dfrac{\rho_{yx}(x_a - x_b)}{\sqrt{2(1 - \rho_{yx}^2)}}$.

Figure 1 illustrates probabilities for different levels of correlation between Y and X and different sizes of differences between X_a and X_b that are denoted D. As can be seen, when both ρ_{yx} and d are small (the lower left corner of Figure 1), the probability of correctly identifying the correct alternative is close to 0.50, i.e., no better than random choice. However, as both the correlation and difference increase in size, the probability of correct choice also increases.

Insert Figure 1 about here

Above we considered probabilities associated with particular observations, i.e., that A is larger than B given that a specific value, x_a, exceeds a specific value, x_b. However, it is also instructive to consider the overall expected accuracy of SV in a given environment or population. To do this, it is necessary to consider the cases where both $X_a > X_b$ and $X_b > X_a$ such that the overall probability is given by $P\{(Y_a > Y_b | X_a > X_b) \cup (Y_b > Y_a | X_b > X_a)\}$ which, since both its components are equal, can be simplified as

$$P\{(Y_a > Y_b | X_a > X_b) \cup (Y_b > Y_a | X_b > X_a)\} = 2 P(Y_a > Y_b | X_a > X_b) \tag{7}$$

To derive analytically the overall probability of correct choice by SV, consider first the integration of $P(Y_a > Y_b | X_a > X_b)$ across all possible values that can be taken by $D = X_a - X_b > 0$. Second, since D is normally distributed with mean of zero and variance of two, the overall probability of success (7) can be re-expressed as:

$$2 \int_0^\infty \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{d^2}{4}\right) \left[\int_{-\infty}^{d*} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \right] dd \tag{8}$$

where $d* = \dfrac{d\rho_{yx}}{\sqrt{2(1-\rho_{yx}{}^2)}}$.

Equal weighting (EW) and multiple regression (MR)

What are the predictive accuracies of models that make use of several, k, cues or variables, $k > 1$? We consider two models that have often been used in the literature. One is equal weighting (EW – see Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). The other is multiple regression (MR). To analyze these models, assume that the criterion variable, Y, can be expressed as a function

$$Y = f(X_1, X_2, ..., X_k) \qquad\qquad (9)$$

where the k explanatory variables are multivariate normal, each with mean of 0 and standard deviation of 1. For EW, the predicted Y value associated with any vector of observed x's is equal to $\frac{1}{k}\sum_{j=1}^{k} x_j$ or $\bar{x}$ where it is assumed that all variables have been scaled to have positive correlations with the criterion.[4] Similarly, the analogous prediction in MR is given by $\sum_{j=1}^{k} b_j x_j$ or $\hat{y}$ where the b_j's are estimated regression coefficients. In using these models, therefore, the decision rules are to choose A over B if $\bar{x}_a > \bar{x}_b$ for EW and if $\hat{y}_a > \hat{y}_b$ for MR (and vice-versa).

How well are EW and MR likely to choose correctly between A and B? Following the same rationale as the single variable (SV) case, we show in Table 1 the formulas used in deriving the analogous probabilities for EW and MR. (We also repeat those for SV.) These are the initial equations (corresponding to equations 2 and 3), the accuracy conditions (corresponding to equation 5), the relevant error variances, and finally s^* used to calculate probabilities – labeled SVs^*, EWs^*, and MRs^*, respectively.

Insert Table 1 about here

As might be expected, all the formulas contain similar terms involving both the size of the difference between observations characterizing A and B, and the extent to which the criterion variable, Y, can be predicted on the basis of the explanatory variables, the X's. There are, however, differences.

---

[4] We return to discuss this point when presenting the simulation below.

For SV, the key factor for correct prediction across a set of cases, with different d values, is the size of ρ_{yx} (as illustrated in Figure 1) – the larger this correlation, the more accurate SV.

In EW, the size of s^* (and hence the probability of correctly discriminating between A and B), is a decreasing function of the average intercorrelation between the X variables – see the formulas in Notes 1 and 2. Holding other values constant, the correlation between Y and $\overline{X}$ increases with k, the number of explanatory variables. However, there are decreasing returns. Whether a variable adds to predictability depends on how it affects both the average correlation with the criterion, i.e., $\overline{\rho}_{yx}$, and the level of average intercorrelation among the predictors, i.e., $\overline{\rho}_{x_ix_j}$ (for a complete analysis see Hogarth, 1978. For empirical evidence, see Ashton, 1986).

As to MR, it is well known that intercorrelation among predictors reduces predictability relative to situations where the explanatory variables are orthogonal. In addition, and particularly when the ratio of observations (n) to explanatory variables (k) is small, MR tends to "overfit" in estimating regression coefficients such that predictability falls short of the R^2 achieved on estimation. For this reason, the error variance and s^* for MR have been derived using R^2_{adj} (i.e., adjusted R^2) instead of R^2 – see Note 3.

The term s^* in Table 1 can be used in different ways. One is to calculate the probabilities that, given differences on the X variables, the models will discriminate between A and B as shown for SV in Figure 1. Another is to predict which model is more likely to make the correct discrimination by comparing the sizes of their s^* statistics. As an example, imagine having to choose between SV and EW. Clearly, SV will have a higher probability of correct prediction than EW if $SVs^* > EWs^*$.

Conceptually, one can think of each s^* as being a ratio of "predicted difference" or *pd* over "prediction error" or *pe*. Thus, SV should be preferred over EW if

$$\frac{pd(SV)}{pd(EW)} > \frac{pe(SV)}{pe(EW)} \tag{10}$$

in other words, if the ratio of "predicted differences" of SV to EW exceeds the ratio of "predicted errors" of SV to EW. Note that this approach neatly takes into account both the actual values of the explanatory variables on which predictions are made in particular cases and the general predictive accuracies of the models. Moreover, the logic used in developing inequality (10) can be extended to all pair-wise comparisons of the three models.

In Table 2, we compare the relative predictive abilities of the different models: SV vs. EW; SV vs. MR; and MR vs. EW. For each comparison, we state, first, the condition (e.g., for SV vs. EW that $\rho_{yx_1} > \rho_{y\bar{x}}$ where we have expanded the latter term using Note 1 of Table 1).[5] Next, using definitions and rearranging terms we re-express the conditions to highlight different factors.

Insert Table 2 about here

Inequality (2.2) shows that, *ceteris paribus*, SV will be more effective than EW when: (1) the ratio of the single variable's cue validity (i.e., ρ_{yx_1}) to the sum of all cue validities is large. Assuming that ρ_{yx_1} is the largest of the cue validities, this will occur when cue validities vary in size; and (2) the average intercorrelation between the explanatory variables (i.e., $\bar{\rho}_{x_i x_j}$) is large (and positive).

---

[5] Henceforth we denote by ρ_{yx_1} the correlation between the criterion and cue of the SV model to distinguish this correlation from the correlations of the criterion with the other *x* variables.

11

Inequality (2.4) compares SV with MR and emphasizes that SV will be more predictive when: (1) $\rho^2_{yx_1}$ is, at least, not much smaller than R^2; and (2) *(n-k)* is small relative to *(n-1)*, i.e., the ratio of variables (cues) to observations is large. Two factors play a role in the first condition: variability in cue validities and positive intercorrelation among the explanatory variables. Both are favorable to SV.

Comparing MR with EW in (2.6), it can be shown that MR is more likely to outperform EW when: (1) there is variability in cue validities; and (2) the ratio of observations to cues is large. An alternative and more complete analysis of the comparison can be found in Einhorn and Hogarth (1975).

These model comparisons can be further illuminated by considering the formulas for overall predictive accuracy such as that for SV derived above (equation 9). This and the analogous formulas for EW and MR are presented in Table 3. Furthermore, the four panels of Figure 2 illustrate the theoretical predicted performance of SV and EW across several scenarios involving three explanatory variables. In each panel, expected percentage correct is plotted as a function of ρ_{yx_1} and the predictive ability of the other variables is varied by panel. In panel (a), both X_2 and X_3 have low predictive ability whereas in panel (d) this is much higher. Panels (b) and (c) represent intermediate values. The panels also show the effects of intercorrelation of the explanatory variables – from almost nothing to 0.8.

Insert Table 3 and Figure 2 about here

As can be seen, the expected performance of SV is – of course – unaffected by characteristics of X_2 and X_3. Moreover, and echoing the preceding discussion, whether the expected performance of EW is better than that of SV depends on the

predictive ability of X_2 and X_3 and the level of average intercorrelation among the predictors, i.e., the effects of $\bar{\rho}_{yx}$ and $\bar{\rho}_{x_ix_j}$ (see Note 1 in Table 1).

Illustration

We simulated several datasets to illustrate the theoretical approach taken above. Our procedure involved seven steps.

(1) We generated datasets with multivariate normal distributions where all variables were standardized and we could specify the covariances between all variables. Each dataset contained 20 observations on a criterion variable and three to five associated explanatory variables.

(2) We sampled half of the observations (i.e., 10) from each dataset at random without replacement thereby creating "fitting" and "holdout" samples.

(3) We considered the 45 possible pairings of observations in the fitting samples (i.e., 10(10-1)/2) and, using the formulas in Table 1 to estimate parameters from the samples, we made theoretical predictions for all 45 pairings for all three models, that is, SV, EW, and MR. Averaging across the 45 probabilistic predictions, we established overall theoretical predictions per model per dataset. For several datasets, these theoretical predictions are detailed in lines 1, 3, and 5 of Table 4 – see also below.

Insert Table 4 about here
----

(4) We then tested how well these theoretical predictions matched the actual criterion. For example, if the overall theoretical prediction of a model for a particular dataset was, say, 75%, we would expect to find, across the dataset, that the model had

in fact classified 75% of cases correctly. These tests of "fit" are shown in lines 2, 4, and 6 of Table 4.

(5) Using the parameters estimated from the fitting samples, we made theoretical predictions for the 45 pairings in each of the holdout samples – lines 7, 9, and 11 of Table 4.

(6) We then tested these predictions. Results are shown in lines 8, 10, and 12 of Table 4.

(7) Step (1) – the initial creation of datasets – was done 50 times for each set of factors investigated (see below). For each of the resulting datasets, Steps (2) through (6) were carried out 20 times such that the outcomes presented in lines 1 through 14 of Table 4 represent averages of 1,000 simulations.

Before commenting on Table 4, we draw attention to lines 13 and 14 which report "fitting" and "realization" statistics for DR or DOMRAN, a strategy based simply on dominance. That is, if one alternative dominates the other, it is chosen. If there is no dominance, choice is made at random. We include DOMRAN because, it has proven to be remarkably effective when cues are binary and provides a useful, naïve baseline with which to compare performance of other simple strategies (Hogarth & Karelaia, 2004). Since DOMRAN has no parameters, we simply show how it performed in the fitting (line 13) and holdout (line 14) samples.

In Table 4, we provide results of some of our simulations which we have classified by four cases: A, B, C, and D. In constructing these datasets, we varied factors identified in Table 2 that affect the relative predictive abilities of the models. These are detailed at the foot of Table 4 and are: (1) the ratio *(n-1)/(n-k)* that captures the effects of variables to observations. For cases A and B, the ratio is 1.3 (9/7),

whereas for cases C and D, it is 1.8 (9/5);[6] (2) the R^2 on initial fit for each dataset. Within cases, we have varied R^2 from low (medium) to high, e.g., from 0.4 to 0.8; (3) average intercorrelation of the explanatory variables or cues. These are varied from low (cases A and C) to moderately high (cases B and D); (4) variability of cue validities – also varied from low to high but within all cases. Low (high) variability corresponds approximately to what Martignon and Hoffrage (1999; 2002) call compensatory (non-compensatory) weighting functions and what Payne, Bettman, and Johnson (1993) refer to as low (high) dispersion.

Ideally, one would like to conduct a complete experimental design in which all factors are manipulated systematically. However, covariation matrices impose restrictions such that one cannot always hold some variables constant. Thus, for example, increasing variability of cue validities also implies increasing the size of ρ_{yx_1} compared to the other cue validities.

Parenthetically, we stress that the average intercorrelations of the cues are calculated for variables that are all positively correlated with the criterion. As noted previously, this specification is required by EW (as well as DR). It leads to the criticism, however, that we have reduced the possibility of negatively intercorrelated explanatory variables and, as such, omitted a potentially important segment of the sample space. Fortunately, this is not the case. First, the scaling of variables makes no difference to the predictive abilities of two models, SV and MR. (SV only uses the most valid variable and ignores the others; in MR, R^2 does not depend on how variables have been scaled.) Second, the mean intercorrelations we use are equal to

---

[6] An anonymous reviewer correctly commented that MR should not really be used unless this ratio is much closer to 1. We agree (cf., the analysis of regression versus equal weights in Einhorn & Hogarth, 1975). However, we explicitly chose larger ratios in our simulations because we wished to approximate the conditions of Gigerenzer et al.'s (1999) 20 datasets which were characterized by similar large ratios (mean of 2.0, median of 1.4).

the average of absolute intercorrelations (i.e., independent of sign) before any rescaling. Thus, the implication of using mean intercorrelations (after rescaling) is that we are effectively sampling situations where, prior to rescaling, inter-cue correlations were both positive and negative.

Consider, first, lines 1 through 6 of Table 4 that deal only with theoretical predictions of percentage correct classifications and their match with the fitting samples. (Note, in making the theoretical predictions for MR here, we have used R^2 and not R^2_{adj}.) For all three models, there is an almost perfect match between theory and fit. Second, examine the data for the holdout samples – lines 7 through 12. For SV and EW, there are almost no differences between theoretical predictions (using fitting sample parameters) and realizations in the holdout samples. For MR, on the other hand, predictions are systematically higher than realizations even though we used R^2_{adj} in our theoretical calculations.[7] This is especially the case for $k = 5$ (i.e., C and D) and reflects the common experience of "overfitting" by regression models – a situation for which the adjustment by R^2_{adj} is inadequate – see footnote 5. (Bold figures in Table 4, e.g., **66**, are used to indicate the largest figure in relevant comparisons, i.e., which model has the best performance on "prediction" or "realization" for particular datasets.)

To illuminate relative model performance in the holdout samples, we provide graphs in the four panels of Figure 3 of trends that are difficult to discern from lines 8, 10, and 12 of Table 4. In Case A, where $\bar{\rho}_{x_i x_j} = 0$, all three models have similar performance when variability in cue validities is low ([max – min] < 0.5). However,

---

[7] Additional and illuminating comparisons are between lines 1 and 8, 3 and 10, and 5 and 12 (theoretical calculations based on statistics observed in the fitting samples versus realizations in the holdout samples). For SV and EW, the matches between predictions and realizations are almost perfect. MR, however, exhibits considerable "shrinkage."

as variability increases, EW performs less well in a relative sense. In Case B, $\bar{\rho}_{x_i x_j} =$ 0.5, and this affects MR negatively relative to Case A. SV, on the other hand, dominates all other models under the conditions illustrated here. Comparing Cases A and B, note also how high intercorrelation among predictors has a negative impact on EW.

Insert Figure 3 about here

Case C illustrates conditions where SV is dominated by both MR and EW, i.e., low $\bar{\rho}_{x_i x_j}$ (= 0.1); high R^2 on initial fit (0.5 to 0.9); and low variability of cue validities ([max − min] ≤ 0.5). Note too that in this panel EW generally performs better than MR. Here the latter is penalized by the number of variables relative to observations. Finally, the relative efficacy of SV is demonstrated in Case D where SV predicts as well or better than the other models. In these conditions, $\bar{\rho}_{x_i x_j}$ is high (= 0.6) − compare with Case B.

Discussion

Throughout this paper we assumed normally distributed variables. However, it should be noted from the derivations for SV in equations (2) through (5) that we do not need to assume that Y and X are normally distributed. The only normality assumptions required concern the error terms, ε_a and ε_b. This also applies to the EW and MR models (see Table 1). Thus, the results presented here are more general than might at first appear. On the other hand, our equations for the overall predictive abilities of the models (see equation 8 and Table 3) do require the normality

assumption and it is for this reason that we have put less weight on these derivations in our simulations.

Predictive models can have many criteria. In this work, we have limited attention to percentage correct because this criterion has been extensively used in binary choice (cf., Gigerenzer et al., 1999). In addition, since the properties of SV, MR, and EW are well-established for squared-error loss functions, the use of percentage correct represents a methodological innovation. Similarly, one can think of extending the current work to situations where choice involves many alternatives (greater than two) and, indeed, we have already made progress on this extension (Hogarth & Karelaia, in preparation).

We return to our starting point. Gigerenzer et al. (1999) found, across 20 datasets, that – with data having "exact quantitative values" – their TTB model and MR shared the same level of cross-validated predictive accuracy of 76% in binary choice. In light of the theoretical results provided above, is this surprising?

First, note that our SV model is not quite the same as TTB in that the latter has a mechanism to resolve "ties" when $x_a = x_b$. In other words, SV is more "frugal" than TTB. And yet, as illustrated by Table 4 and Figure 3, we can predict and observe situations where the predictive accuracy of SV is greater than that of MR.[8] As observed, the main factors that determine the relative predictive accuracies of SV and MR center on the level of intercorrelations between predictors and the ratio of predictors to observations.[9] Clearly, one can define environments that are more or less "friendly" to SV and MR (cf., Payne et al., 1993; Shanteau & Stewart, 2000).

---

[8] We can, of course, also predict when SV will perform at the same level or worse than MR.
[9] Unfortunately, we do not have access to the quantitative version of the data of Gigerenzer et al. (1999) and thus cannot estimate the relevant predictive factors. The binary cue versions of the datasets are available at the website, http://www-abc.mpib-berlin.mpg.de/sim/Heuristica/environments/

The comparisons made above between SV and MR involved averaging the results of all possible pairings of observations in the fitting and holdout samples (i.e., 45 pairs on each occasion). However, a person choosing between SV and MR on a particular occasion would not be obliged to choose the model that was generally better (as also formalized by comparing equation 8 for SV with its analog for MR in Table 3). Instead, as indicated by inequality (10) and the equations in Table 1, predictions of which model to use for particular cases can be tailored to the characteristics of each case. For example, conditional on a particular observation, one model could predict, say, alternative A with probability 0.7 whereas the other could predict alternative B with, say, probability 0.6. In this case, the better solution would be to use the model with the greater probability of being correct even if, in general, the factors outlined above favor the other model.

This paper adds to the growing literature suggesting that, in many cases, "less is more" (see, e.g., Hertwig & Todd, 2003). But to what extent are these phenomena unusual? One way of conceptualizing this issue is to recall the insights of Coombs and Avrunin (1977) concerning why certain functions are single-peaked across dimensions such as time or amount of information. Simply put, assume we are dealing with phenomena that are a mix of both "good" and "bad" dimensions (e.g., information and cost). Furthermore, assume that "goods satiate" whereas "bads escalate" in either physical or psychological terms. The net effect is a single-peaked function over the dimension of interest. Thus, it can be argued, many phenomena in the environment do produce "less is more" effects over relevant dimensions (e.g., time, information, etc). Our purpose in this paper, however, is not to advocate the use of single or multiple variables in decision making. It is to illuminate the conditions under which one or the other is preferable.

Finally, is it possible to reconcile the apparent efficacy of single-variable models (or "one reason decision making") with common sense? Perhaps it is a case of heeding the often repeated injunction – "Keep your eye on the ball!" – but also knowing when this is appropriate.

References

Ashton, R. H., (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes*, 38 (3), 405-414.

Coombs, C. H., & Avrunin, G. S. (1977). Single-peaked functions and the theory of preference. *Psychological Review*, 84 (2), 216-230.

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

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

Fasolo, B., McClelland, G. H., & Todd, P. M. (2004). *Escaping the tyranny of choice: When fewer attributes make choice easier*. Working paper, Max-Planck Institute for Human Development, Berlin.

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

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

Hertwig, R., & Todd, P. M. (2003). More is not always better: The benefits of cognitive limits. In D. J. Hardman & L. Macchi (eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making*. Chichester, England: Wiley.

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

Hogarth, R. M., & Karelaia, N. (2004). *"Take-the-Best" and other simple strategies: Why and when they work "well" in binary choice*. Barcelona: Universitat Pompeu Fabra, DEE working paper.

Hogarth, R. M., & Karelaia, N. (in preparation). Simple models for multi-attribute choice with two or more alternatives: Identifying differential model performance. Barcelona: Universitat Pompeu Fabra.

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

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G Gigerenzer, P. M. Todd and 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.

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

Shanteau, J., & Thomas, R. P. (2000). Fast and frugal heuristics: What about unfriendly environments? *Behavioral and Brain Sciences*, 23, 762-763.

<div align="center">Table 1 – Key formulas for different models</div>

Single variable (SV)	Accuracy condition	Error variance	$\underline{s^*}$	
$Y_a = \rho_{yx} X_a + \varepsilon_a$	$\rho_{yx}(X_a - X_b) > \varepsilon_b - \varepsilon_a$	$2(1-\rho_{yx}^2)$	$\dfrac{\rho_{yx}(x_a - x_b)}{\sqrt{2(1-\rho_{yx}^2)}}$	(SVs^*)
$Y_b = \rho_{yx} X_b + \varepsilon_b$				

Equal weights (EW)

$Y_a = (\rho_{y\bar{x}}/\sigma_{\bar{x}})\bar{X}_a + v_a$	$(\rho_{y\bar{x}}/\sigma_{\bar{x}})(\bar{X}_a - \bar{X}_b) > v_b - v_a$	$2(1-\rho_{y\bar{x}}^2)$	$\dfrac{(\rho_{y\bar{x}}/\sigma_{\bar{x}})(\bar{x}_a - \bar{x}_b)}{\sqrt{2(1-\rho_{y\bar{x}}^2)}}$	(EWs^*)
$Y_b = (\rho_{y\bar{x}}/\sigma_{\bar{x}})\bar{X}_b + v_b$				

Multiple regression (MR)

$Y_a = \hat{Y}_a + \omega_a$	$(\hat{Y}_a - \hat{Y}_b) > \omega_b - \omega_a$	$2(1-R_{adj}^2)$	$\dfrac{(\hat{y}_a - \hat{y}_b)}{\sqrt{2(1-R_{adj}^2)}}$	(MRs^*)
$Y_b = \hat{Y}_b + \omega_b$				

Notes: 1. $\rho_{y\bar{x}} = \bar{\rho}_{yx}\sqrt{\dfrac{k}{1+(k-1)\bar{\rho}_{x_ix_j}}}$ where k = number of x variables, $\bar{\rho}_{yx}$ = average correlation between Y and the X's, and $\bar{\rho}_{x_ix_j}$ = average intercorrelations amongst the X's.

2. $\sigma_{\bar{x}} = \sqrt{\dfrac{1}{k}\left(1+(k-1)\bar{\rho}_{x_ix_j}\right)}$

3. $R_{adj}^2 = 1-\left(1-R^2\right)\dfrac{(n-1)}{(n-k)}$ where n = number of observations

Table 2 – Comparison of predictability of models

SV vs. EW

Condition:
$$\rho_{yx_1} > \overline{\rho}_{yx}\sqrt{\frac{k}{1+(k-1)\overline{\rho}_{x_ix_j}}} \qquad (2.1)$$

Can be re-expressed as
$$\frac{\rho_{yx_1}}{\sum_{j=1}^{k}\rho_{yx_j}} > \sqrt{\frac{1}{k\left(1+(k-1)\overline{\rho}_{x_ix_j}\right)}} \qquad (2.2)$$

SV vs. MR

Condition:
$$\rho_{yx_1}^2 > R_{adj}^2 \qquad (2.3)$$

Can be re-expressed as
$$\frac{1-\rho_{yx_1}^2}{1-R^2} < \frac{n-1}{n-k} \qquad (2.4)$$

MR vs. EW[*]

Condition:
$$R_{adj}^2 > \rho_{y\overline{x}}^2 \qquad (2.5)$$

Can be re-expressed as
$$1-\left(1-R^2\right)\frac{(n-1)}{(n-k)} > \rho_{y\overline{x}}^2 \qquad (2.6)$$

---

[*] See also Einhorn and Hogarth (1975).

Table 3 – Formulas for overall predictive accuracy of models

|	Relevant difference (D)	Variance of the difference	Overall expected probability of success	$\underline{d}^*$
SV	$(X_a - X_b)$	2	$2\int\limits_{0}^{\infty}\dfrac{1}{2\sqrt{\pi}}\exp\left(-\dfrac{d^2}{4}\right)\left[\int\limits_{-\infty}^{d^*}\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{z^2}{2}\right)dz\right]dd$	$\dfrac{d\rho_{yx}}{\sqrt{2(1-\rho_{yx}^{2})}}$
EW	$(\overline{X}_a - \overline{X}_b)$	$2\sigma_{\bar{x}}^2$	$2\int\limits_{0}^{\infty}\dfrac{1}{2\sigma_{\bar{x}}\sqrt{\pi}}\exp\left(-\dfrac{d^2}{4\sigma_{\bar{x}}^2}\right)\left[\int\limits_{-\infty}^{d^*}\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{z^2}{2}\right)dz\right]dd$	$\dfrac{d\rho_{y\bar{x}}/\sigma_{\bar{x}}}{\sqrt{2(1-\rho_{y\bar{x}}^{2})}}$
MR	$(\hat{Y}_a - \hat{Y}_b)$	$2R^2$	$2\int\limits_{0}^{\infty}\dfrac{1}{2R\sqrt{\pi}}\exp\left(-\dfrac{d^2}{4R^2}\right)\left[\int\limits_{-\infty}^{d^*}\dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{z^2}{2}\right)dz\right]dd$	$\dfrac{d}{\sqrt{2(1-R^2)}}$

Table 4 -- Model fits and predictions (%)

|		Case A (k=3)					Case B (k=3)					Case C (k=5)					Case D (k=5)				
SV	1. Theory	63	65	72	75	80	66	69	71	76	80	62	66	69	69	73	61	62	71	75	80
	2. Fitting	62	64	71	75	80	65	69	70	76	80	62	65	69	69	72	60	62	70	75	80
EW	3. Theory	66	69	71	73	75	63	63	63	67	68	66	71	73	77	83	61	63	65	68	69
	4. Fitting	64	68	70	73	74	62	63	63	66	67	66	70	72	76	81	61	63	65	68	67
MR	5. Theory[1]	72	75	78	81	85	69	71	72	77	80	75	80	81	85	90	66	68	73	76	80
	6. Fitting	71	74	77	81	85	69	71	71	77	80	75	80	81	84	90	67	69	72	77	80
SV	7. Prediction	62	64	71	74	80	65	68	**71**	**76**	**80**	62	65	68	69	73	60	62	**71**	**75**	**80**
	8. Realization	62	64	71	75	79	**64**	**68**	**71**	**76**	**80**	62	65	68	68	72	**60**	60	**70**	**76**	**80**
EW	9. Prediction	65	68	70	73	75	62	63	63	66	67	66	71	72	77	83	61	63	65	68	69
	10. Realization	**64**	**68**	71	73	74	62	62	64	66	67	**66**	**71**	**71**	**76**	82	**60**	**62**	65	68	67
MR	11. Prediction[2]	**70**	**73**	**77**	**79**	**83**	**68**	**69**	**71**	75	79	**72**	**76**	**77**	**82**	**87**	**63**	**65**	69	73	77
	12. Realization	62	66	**72**	**76**	**80**	63	66	69	75	79	62	69	69	75	**84**	59	61	67	73	78
DR	13. Fitting	56	57	58	59	60	60	60	60	63	64	53	54	54	55	55	57	59	60	61	62
	14. Realization	56	57	59	59	60	59	59	60	63	63	53	54	54	55	55	57	58	60	61	62
	Variability of cue validities: max - min	0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7	0.2	0.3	0.3	0.4	0.5	0.0	0.1	0.3	0.4	0.5
	Average intercorrelation of x's	0.0	0.0	0.0	0.0	0.0	0.5	0.4	0.5	0.5	0.5	0.1	0.1	0.1	0.1	0.1	0.6	0.6	0.6	0.6	0.6
	R^2 (MR) -- fit	0.4	0.5	0.6	0.7	0.8	0.3	0.4	0.4	0.6	0.7	0.5	0.6	0.7	0.8	0.9	0.3	0.4	0.5	0.5	0.7
	(n-1)/(n-k)	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.8

[1] Initial fit used R^2

[2] Prediction made using adjusted R^2

26

Figure 1: Probability of SV making correct prediction/choice as a function of ρ_{yx} and $d\ (= x_a - x_b)$.
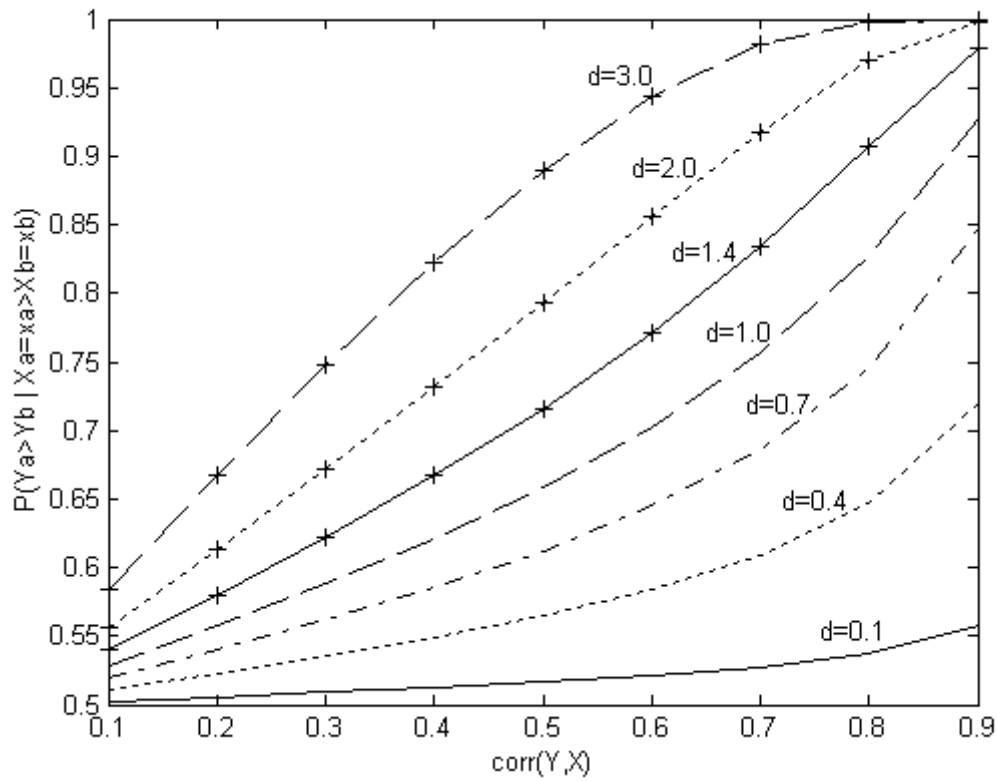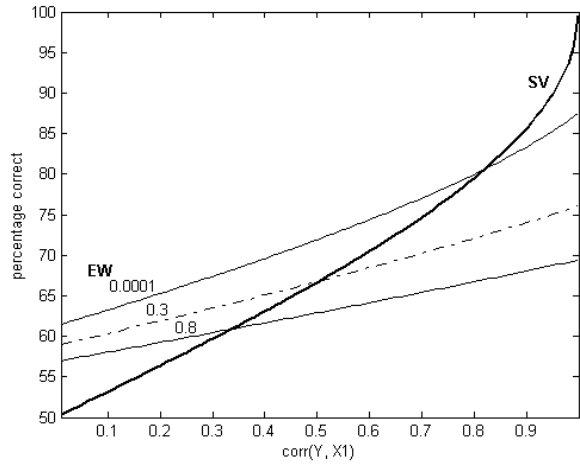
Figure 2: The theoretical performance of SV and EW (percentage correct predictions/choices) as a function of ρ_{yx_1} in 3-cue environments.
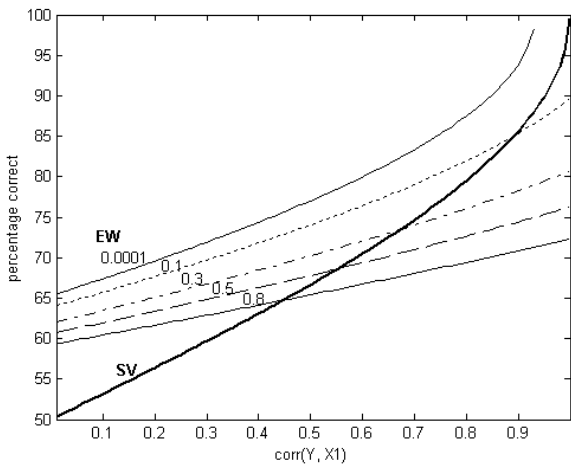
a) corr(Y, X_2)=0.1, corr(Y, X_3)=0.1
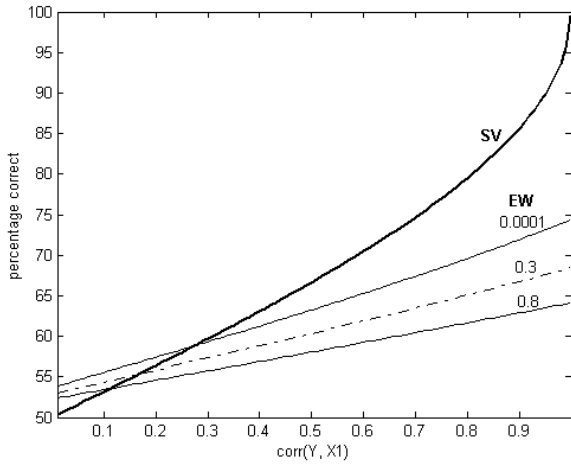
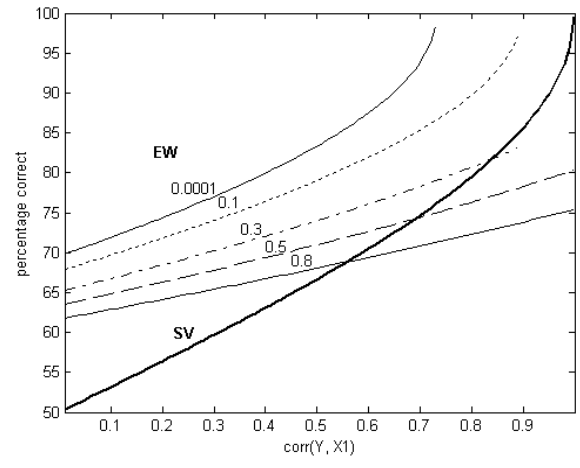b) corr(Y, X_2)=0.5, corr(Y, X_3)=0.1

c) corr(Y, X_2)=0.5, corr(Y, X_3)=0.3

d) corr(Y, X_2)=0.5, corr(Y, X_3)=0.5

The performance of EW is represented for different levels of $\overline{\rho}_{x_i x_j}$.

Figure 3: Percentage correct predictions by SV, MR, EW, and DR for conditions specified in Table 4 (cases A, B, C, and D).