Robust decision making in a non-linear world

Michael R. Dougherty

University of Maryland

&

Rick P. Thomas

University of Oklahoma

Corresponding author:

Michael Dougherty

Department of Psychology

University of Maryland

College Park, MD 20782

mdougher@umd.edu

Abstract

The authors propose a general modeling framework called the General Monotone Model (GeMM), which allows one to model psychological phenomena that manifest as non-linear relations in behavior data without the need for making (overly) precise assumptions about functional form. Using both simulated and real data, the authors illustrate that GeMM performs as well or better than standard statistical approaches (including ordinary least-squares, robust, and Bayesian regression) in terms of power and predictive accuracy when the functional relations are strictly linear, but outperforms these approaches under conditions in which the functional relations are monotone, but non-linear. Finally, the authors recast their framework within the context of contemporary models of behavioral decision making, including the lens model and the take-the-best heuristic, and use GeMM to highlight several important issues within the judgment and decision making literature.

Key words: prediction, heuristics, choice, take-the-best, regression, monotone

The nature of the phenomena typically studied in the behavioral and social sciences rarely reflect simple input-output functions; rather, they are often an emergent property of complex social, cognitive, and neural systems. Constructs such as intelligence, anxiety, prejudice and depression no-doubt correspond to some fundamental property of the human condition, yet they also have no tangible corresponding representation in the physical world; they are realized only through psychological measurement. The distinction between phenomena and data is subtle, but important. Phenomena are the constructs or entities that theories are meant to explain; data are observables that we use as proxies of the phenomena (Bogen & Woodward, 1988).

The critical link between data and phenomena are the set of techniques used for modeling the data, as these techniques allow one to state facts about the data that are used to support the theory of the phenomena. Given this critical link, any assumptions made in the course of modeling data have implications for our understanding of the corresponding phenomena. Unfortunately, for the purposes of (statistical) modeling scientists often make strong assumptions about functional relationships when describing data, without realizing (or acknowledging) that assumptions about one's data are not mere abstractions but are statements about the nature of the underlying phenomena to which the data correspond (cf. Berk, 2004).

The link between the statistical modeling of data and the construction of theories about phenomena exists for any theory leveraged against real data. While it easy to think of the assumptions embodied in statistical models as applying purely at the level of statistical analysis, these assumptions are often (either implicitly or explicitly) carried forth into theory. At the implicit level, the theoretical conclusions drawn from data are conditioned on the assumptions of the statistical analysis upon which the conclusions were based. Violations of assumptions

undermine statistical conclusions validity, and by extension undermine the process of theory testing.

However, the link is even more explicit in cases in which the statistical modeling framework is used as a description of psychological process. Indeed, it is commonplace for statistical models to be exapted as paramorphic models of human information processing (Gigerenzer, 1991). For example, the ANOVA framework was used by Norman H. Anderson's Information Integration Theory (1968, 1970, 1981) and Harold H. Kelley's Covariation Theory (1967; 1973) to model impression formation and causal attribution. More recently, Bayesian models utilizing principles from statistics and machine learning have been imported into psychology under the guise of computational theories of mind (Tenenbaum, Griffiths, & Kemp, 2006; Griffiths & Tenenbaum, 2006; and Griffiths, Steyvers, & Tenenbaum, 2007). Of particular relevance to the present article is multiple-linear regression, which has had a long and storied history as a descriptive model of human judgment (Brunswick, 1955; Hammond, 1955; Hursch, Hammond, & Hursch, 1964; Hoffman, 1960; Hogarth & Karalaia, 2007), a prescriptive model for clinical judgment (Meehl, 1954; Dawes & Corrigan, 1974; Dawes, Faust & Meehl, 1989), and as a normative benchmark against which to compare heuristic models (Karelaia & Hogarth, 2008; Payne, Bettman & Johnson, 1993).

In terms of description, the linear model has been the "work horse of judgment and decision making research" (pp. 734 Hogarth & Karelaia, 2007) for over 60 years. It is also intimately tied to the evaluation and modeling of heuristic mechanisms put forth by Gigerenzer and colleagues (Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & the ABC group, 1999). Clearly, the assumptions embodied in the aforementioned statistical models carry-forth into the corresponding theoretical frameworks. Thus, assumptions made for descriptive purposes, such as

the assumption of linearity, clearly imply that the explanation of the phenomena requires that assumption. However, what if the data, or the underlying phenomena, are not linear?

The overarching goal of the present paper is to explore the costs of assuming linearity when linearity does not hold, and to propose an alternative approach to statistical decision making and behavioral decision making that does not require strong assumptions about ones data. Our approach relaxes assumptions about functional form, and therefore allows one to model any relationship (linear or non-linear) characterized by monotonicity. In what follows, we first review evidence that many psychological phenomena are characterized by non-linearity's. We then describe a general modeling framework, dubbed the General Monotone Model (GeMM), which allows one to model monotone relations without the need for making precise assumptions about functional form. Using both simulated and real data, we illustrate that GeMM performs as well or better than standard statistical approaches (including ordinary least-squares, robust, and Bayesian regression) in terms of power and predictive accuracy when the nature of data is linear, but outperforms these approaches under conditions in which statistical relationships are monotone, but non-linear. Finally, we recast our framework within the context of contemporary models of behavioral decision making, including the lens model and the take-the-best heuristic, and use GeMM to highlight important issues within the judgment and decision making literature.

Non-linearities in Psychological Research

A core assumption of our work is that the functional relationships that describe data within the behavioral and social sciences are often non-linear. There are many examples of this within psychological science. For instance, many psychophysical judgments are approximated by a logarithmic function (e.g., the Weber-Fechner law). The relationship between practice and skilled performance is approximated by a power function (e.g., the power law of practice, Newell

& Rosenbloom, 1981; but see Heathcote, Brown, & Mewhort, 2002). Cumulative retrieval functions in verbal-fluency tasks follow a negatively accelerated exponential function (Bousfield & Sedgwick, 1944). In economics, subjective expected utility is non-linearly related to value, with the form of the nonlinearity differing for gains, a concave function versus losses, a convex function (Kahneman & Tversky, 1979). In the decision-making literature, the relationship between perceptions of probability and objective probabilities follows a sigmoidal function (DuCharme, 1970; Erev, Wallsten, & Budescu, 1994), whereas the relationship between frequency estimates and actual frequency is sometimes exponential (Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). Finally, children (Thompson, & Opfer, 2008) and adult (Longo, & Lourenco, 2007) perceptions of numbers are systematically distorted, such that differences between larger values are compressed relative to smaller values, roughly following a logarithmic function. The above examples illustrate the wide variety of data patterns observed in psychological science, and therefore serve to ground the central thesis of the work presented here, namely, that the behavioral responses researchers often measure are frequently non-linear. [1]

To be sure, there is good-evidence that phenomena within psychological science, or at least our measurement of them, are often non-linear. Yet the linear model continues to be one of the most widely used approaches to both the analysis of data and the construction of theories about phenomena. One important cost of making strong assumptions about the nature of the data is that one risks missing important statistical relationships that may exist, but which the linear model is ill equipped to detect. However, a far more important cost is that one could clearly misrepresent

[1] While in some cases the descriptive label afforded by specifying the particular functional form may be useful for theoretical purposes (e.g., identifying laws of practice or psychophysics), in the vast majority of cases it is unnecessary for descriptive purposes. Indeed, the examples provided above merely represent a sampling of applications in the behavioral sciences where work has been done to identify the function that best approximates the data.

the correspondence between the data and the phenomena of interest. As Berk (2004) notes, "statistical assumptions invoked to justify an analysis of real data are not just statistical abstractions but statements about nature. If the assumptions made so that a statistical analysis plays through are factual nonsense, one surely has to wonder about the conclusions that follow (pg. 2)." Put another way, by making rigid assumptions in the modeling of data, one implicitly endorses those assumptions as a description of the underlying phenomena. The conclusions drawn from any statistical analysis is conditional on having satisfied a set of assumptions, and by extension, so too are any theoretical statements regarding the corresponding phenomena based on those data.

If the functions that best describe behavioral regularities are non-linear, then the question is how best to model such phenomena. One approach, of course, is to be as precise as possible when making assumptions about data and to fit the function directly, for example by fitting non-linear regression models that estimate particular non-linear functions (e.g., Young, De Leeuw, & Takane, 1976). This approach is tantamount to endorsing more stringent assumptions about the data. The benefit of this approach is that one may gain a deeper understanding of the phenomena of interest. There are drawbacks, however, as one risks overfitting the data, and losing the ability to predict new observations accurately, or worse, mischaracterizing the underlying phenomena (MacCallum, Cornelius, & Champney, 1979). Moreover, in many cases it is even unreasonable to assume that the behavioral responses observed through psychological measurement necessarily are linearly related to the underlying psychological variable.

Partly because of the problems with modeling via curve fitting, a far more common approach to modeling psychological data involves applying transformations to the data or ignoring the potential for non-linearity's, and modeling the data *as-if* it were linear. This approach is

evidenced by the widespread use of the linear model in the analysis of data. However, the linear model still suffers from the same problems as does modeling particular functional relationships; only in this case scientists restrict themselves to the assumption of linearity rather than one or another functional form. Obviously, there may be cases in which both the data and the corresponding underlying phenomena are linear, in which case the assumption of linearity is entirely justified both in the analysis of the data and in describing the underlying phenomena. However, in a vast majority of cases, researchers do not have privileged access to information beyond the sample at hand that allows them to validate whether particular assumptions such as linearity apply at the level of the population, or whether it is characteristic of the underlying phenomena. Nevertheless, because the linear model has been shown to be relatively robust to non-linearity's and violations of assumptions, it is often used by default (c.f., Cliff, 1996).

Unfortunately, data in the psychological sciences rarely affords one with the luxury of making strong inferences about functional form, linear or otherwise. Making matters more difficult is the fact that "few social scientific theories offer any guidance as to function form whatsoever. Statements like y increases with X (monotonicity)" (Beck & Jackman, 1998, pp 597) typify the level of specificity offered by most theories (see also Cliff, 1993). Thus, one often can't even fall back on theory as a means of justifying particular assumptions.

In this paper, we advocate a third approach to modeling behavioral data -- one that involves making *less restrictive assumptions about the data* -- and which is appropriate for modeling any behavioral regularity characterized by monotonicity. The advantage of our proposed approach is that it entails making fewer assumptions, so there is less room for one to violate the assumptions. As we show below, relaxing assumptions about functional form and modeling data at the level of monotonicity can yield better statistical power than models that assume linearity. Moreover,

inasmuch as statistical assumptions carry-forth into ones theory of the phenomena of interest, statistical algorithms that make fewer assumptions should be preferred as a matter of parsimony. In what follows, we outline one such algorithm that follows this third approach.

The General Monotone Model

The fundamental goal within GeMM is to find the coefficients (parameter weights) that minimize the lack of fit between the predicted values of Y and the observed values. However, unlike least-squares regression procedures, which minimize the squared deviations between the predicted and observed values and solve for coefficients that yield the best *linear* solution, GeMM minimizes the number of incorrectly predicted paired comparisons and solves for the coefficients that yield the best *monotonic* (i.e., rank order) solution.

The power to resolve rank order from paired comparisons stems from the fact that the number of constraints on the ordered relationship grows exponentially as a function of N, as given by the fact that there are $N(N\text{-}1)/2$ paired comparisons. As pointed out by Shepard (1962; 1966) in his development of Ordinal Multidimensional Scaling, these ordinal constraints also enable one to recover close approximations of the metric (i.e., least-squares) solutions from data without computing squared distances. Within the context of GeMM, the constraints forced on the procedure by using paired comparisons enable the model to recover close approximations of the metric regression coefficients, without using least-squares estimation procedures. The advantage of using paired comparisons as opposed to least-squares is that one need not specify the form of the relationship a priori, and there is no need to apply (monotone) transformations to the data to make it fit the assumptions of the model (Cliff, 1993). Moreover, GeMM's estimated coefficients are robust to the presence of outliers and the marginal distribution of the criterion variable, because the rank order ignores the precise distances between points on the criterion. The results

of our competitive model tests suggest that GeMM is a stable algorithm that predicts well in a variety of contexts.

The basic framework for implementing GeMM involves 3 steps:

(1) Identify the set of potential models, with each model representing a collection of variables to be evaluated, or which are theoretically meaningful. In the most general case, one might be interested in evaluating all possible subsets (i.e., models) of $k \geq 1$ variables.

(2) Estimate the parameter values for each model that maximizes the correspondence between the predicted values and the observed values. In GeMM, the goal is to find parameter coefficients that minimize the lack of rank-order correspondence between the predicted and actual values, in terms of the underlying dominance structure.

(3) Correct for model complexity. For the purposes here, we have implemented GeMM within the context of the Bayesian Information Criterion (BIC) framework, though other definitions of model complexity can be readily applied.

In its simplest form, GeMM consists of a one-parameter model (i.e., one predictor), which is used to predict the criterion of interest. In this context, GeMM is actually identical to the well-known measure of association, tau, specified by Kendall (1938). Thus, to ease the introduction of GeMM, we build the case for GeMM by drawing parallels with Kendall's tau, and then extend its application to the multiple predictor case.

To start, assume one wants to know the strength of relationship between two variables, X (e.g., height) and Y (e.g., weight). Tau is a valid measure of the strength of any monotonic relationship, irrespective of the measurement properties of the data, and does not require strong assumptions about functional form (Gonzalez & Nelson, 1996).

Tau estimates the degree of monotonic relationship by counting the number of paired comparisons for which the ordinal relationship amongst each pair of *X*'s is concordant (i.e., in agreement) or disconcordant (i.e., in disagreement) with its corresponding pair of *Y*'s. For example, in estimating the relationship between height and weight, one need only count the number of pairs of individuals in which the taller person is heavier than the shorter person, and the number of pairs of individuals in which the shorter person is heavier than the taller person. If there are 10 individuals in the sample, there are 10(10-1)/2 = 45 such comparisons. Formally, tau is defined as,

$$\tau(X,Y) = (C-D) / \text{sqrt}[(\text{Pairs} - T_p)*(\text{Pairs} - T_c)] \text{ where} \quad (1)$$

$$C = Prop(Y_i > Y_j \cap X_i > X_j) + Prop(Y_i < Y_j \cap X_i < X_j) \quad (2)$$

$$D = Prop(Y_i > Y_j \cap X_i < X_j) + Prop(Y_i < Y_j \cap X_i > X_j) \quad (3)$$

$$T_p = Prop(Y_i \geq Y_j \cap X_i = X_j) + Prop(Y_i \leq Y_j \cap X_i = X_j) \quad (4)$$

$$T_c = Prop(Y_i = Y_j \cap X_i \leq X_j) + Prop(Y_i = Y_j \cap X_i \geq X_j) \quad (5)$$

$$\text{Pairs} = N(N-1)/2 \quad (6)$$

In words, *C* is the proportion of paired comparisons in which the values of *X* and *Y* are in concordance (i.e., the ordinal relationship between Y_i and Y_j matches the ordinal relationship between X_i and X_j). *D* is the proportion of paired comparisons in which the values of *X* and *Y* are in disconcordance (i.e., the ordinal relationship between Y_i and Y_j is opposite the ordinal relationship between X_i and X_j). T_p is the proportion of paired comparisons in which the predictor variable is tied and T_c is the proportion of paired comparisons in which the criterion variable is tied. P is the total number of paired comparisons.

Equation 1 can be rescaled to estimate the Pearson product moment correlation, r. Following Kendall (1970),

$$r_\tau \approx \sin(\pi/2 \bullet \tau) \quad (7)$$

where r_τ is the value of r estimated from tau. Although Kendall's tau to r transformation is not an unbiased estimate of r, it is a close approximation even at modest sample sizes (N = 50; Rupinski & Dunlap, 1996). Importantly, prior work has shown that the standard error of r_τ is greater than the corresponding standard error of r calculated on the same samples (Rupinski & Dunlap, 1996), but where the variance of both estimates shrink with increases in N. We return to this issue later.

The advantage of using tau to model statistical relationships is that it is invariant to monotone transformation. Thus, tau is robust, in the sense that within the class of monotone relationships, it is insensitive to the particular functional form. Moreover, the use of equation 7 to rescale tau provides a means of estimating the value of r under any order preserving (monotone) transformation that linearizes the data, *without the need to actually apply the transformation.* The advantage of this approach is that it eliminates the need for the researcher to explore the data to find the one transformation that enables the data to meet the assumptions of the statistical test. Moreover, it also removes any subjectivity, and for that matter implicit bias, toward choosing to transform the data or not, and if so which transformation to choose.

A model-based approach to monotonic association

Much like one can express the linear relationship between *X* and *Y* with a least-square prediction model, one can express the monotonic relationship between *X* and *Y* in terms of an analogous model, as defined by

$$\hat{Y} = \beta X \quad (8)$$

In equation 6, one wishes to find a value for β that minimizes the *incorrectly* predicted paired comparisons. We substitute $\hat{Y}$ for *X* to reflect this subtle change:

$$\tau(\hat{Y}, Y) = (C\text{-}D) / \text{sqrt}[(\text{Pairs} - T_p)*(\text{Pairs} - T_c)] \quad (9)$$

$$C = Prop(Y_i > Y_j \cap \hat{Y}_i > \hat{Y}_j) + Prop(Y_i < Y_j \cap \hat{Y}_i < \hat{Y}_j) \quad (10)$$

$$D = Prop(Y_i > Y_j \cap \hat{Y}_i < \hat{Y}_j) + Prop(Y_i < Y_j \cap \hat{Y}_i > \hat{Y}_j) \quad (11)$$

$$T_p = Prop(Y_i \geq Y_j \cap \hat{Y}_i = \hat{Y}_j) + Prop(Y_i \leq Y_j \cap \hat{Y}_i = \hat{Y}_j) \quad (12)$$

$$\mathrm{T_c} = Prop(Y_i = Y_j \cap \hat{Y}_i \leq \hat{Y}_j) + Prop(Y_i = Y_j \cap \hat{Y}_i \geq \hat{Y}_j) \quad (13)$$

With one predictor, only the sign of β matters, which reflects the direction of the relationship of $\hat{Y}$ and Y. Thus, within the context of simple monotonic prediction (equation 8), the goal is to assess the degree of monotonic association between Y and its predicted value ($\hat{Y}$). Unlike predicting metric values, no intercept is necessary because adding a constant to $\hat{Y}$ (in equation 8) would not affect equations 9-13.

For convenience, it is easiest to think of the β's for the class of one-parameter models as being constrained to {+1,0,-1}, such that the relationship between Y and its predicted value ($\hat{Y}$) is always expressed as a positive value of tau, but where the beta coefficient defines the relationship. The simple model in (8) can be generalized to include multiple predictors:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (14)$$

In this case, the estimated coefficients allow the variables to differentially contribute to the prediction equation to determine the model that maximizes the rank order concurrence between Y and $\hat{Y}$.

Equations 9 through 14 provide the computational basis for computing the degree of monotonic relationship between an additive combination of the predictors (the X's) and a criterion variable (Y). Thus, 1 – *tau* provides the quantity that should be minimized and is at its maximum when prediction is at chance and is zero when prediction is perfect (i.e., *tau* = 1).

Generating Weights

Parameter estimation in GeMM requires searching the parameter space to identify coefficients that maximize the value of tau. Given that the search space grows exponentially with the number of predictors, the parameter space to be searched is extremely large even with a small number of predictors. For example, with $k = 9$ predictors, there are $2^{k=9} = 512$ possible models, excluding interaction terms, with each parameter requiring estimation in metric space. Fortunately, a number of search algorithms have been developed that make searching complex parameter spaces fairly straightforward, including genetic algorithms, simulated annealing, and memetic algorithms, among others. Thus far, we have used genetic algorithms (GA), which are based on the principles of natural selection, and which are particularly effective in large and complex search spaces (Goldberg, 2002). Genetic algorithms consist of several steps: (1) generate a random population of parameter vectors, (2) evaluate the fitness of each member of the population, (3) identify those models that are 'best' and carry them over to the next population (i.e., elitism), (4) stochastically sample models (i.e., parents) from the population in the first step according to their fitness, and create a new population of models (i.e., offspring) by directly copying the sampled parents or probabilistically recombining the sampled models, and then probabilistically add "noise" to the parameters (i.e., mutation) before adding the offspring to the new population and (5) return to step 2. This process is implemented iteratively until the algorithm converges on a solution. Genetic algorithms are known to function well in complex search spaces, though have a tendency to find local minima in some search landscapes. Local minima can be cross-checked by generating new random starting populations through repeated runs.

Occams Razor: Trading fit for complexity

An important consideration in model building is model complexity: one often wishes to find the best fitting statistical model while minimizing the model complexity (see Myung & Pitt, 2004). In standard regression contexts, a number of alternatives have been proposed for trading model complexity for fit, including Akaike's Information Criterion (AIC, Akaike, 1974), Schwarz's Information Criterion (SIC; Schwarz, 1978, also known as the Bayesian Information Criterion, BIC), and the Risk Inflation Criterion (RIC; Foster & George, 1994), amongst others. All of these alternatives have been developed within least-squared or maximum likelihood frameworks, therefore preventing their straightforward application to non-parametric estimators, such as tau. However, the relationship between tau and r stated in equation 5 enables us to work within the context of the non-parametric estimator tau, and still utilize model selection procedures developed for maximum likelihood estimation. In the analyses that follow, we employ a version of the BIC based on R^2 (see Raftery, 1995), though other model selection procedures can be used. Raftery (1995) showed that the BIC could be estimated from

$$\mathrm{BIC} = N \log(1 - R^2) + k \log(N) \qquad (14)$$

where N is the sample size, R^2 is the squared multiple correlation, and k is the number of parameters. Substituting equation 5 for the value of R^2 yields:

$$\mathrm{BIC}_\tau = N \log(1 - (sin[pi/2\,\tau])^2) + k \log(N). \qquad (15)$$

Equation 15 is the value of the BIC estimated from the tau to r transformation. However, because the value of r_τ shows greater variability than r, we use an adjusted form of r_τ based on sample size and the number of predictors used in the regression. Specifically, we define r_τ' as,

$$r_\tau' = sin[pi/2\,\tau\omega], \qquad (16)$$

where

$$\omega = (N-P-1)/N. \qquad (17)$$

ω is a weighting function based on the number predictors, P, used in the regression and sample size, N. ω serves to de-weight the value of tau for smaller sample sizes, and therefore reduces the variance of the tau to r transformation. Because ω goes to 1.0 as N increases, the asymptotic value of the tau to r transformation is preserved.[2] Substituting r_τ' into equation 14 gives us

$$\mathrm{BIC}_\tau' = N \log(1 - r_\tau'^{\,2}) + k \log(N). \qquad (18)$$

Model selection based on equation 18 (BIC_τ') is assessed on the fit of the model to the data as given by the degree of monotonic relationship expressed by the tau to r transformation, adjusted for model complexity. The reliance on r^2_τ' as opposed to r^2 results in a model selection procedure that is invariant to monotone transformation – a property we illustrate next.

Competitive model testing: How does GeMM compare to the linear model?

Equations 9 through 14 provide the computational basis for the GeMM, whereas equation 18 provides a rule for trading complexity for fit. However, how does GeMM perform relative to the linear model? To address this question we conducted a series of modeling competitions and evaluated GeMM's performance on four criteria: (a) the hit rate (i.e., likelihood of detecting true effects), (b) the false-positive rate (i.e., type-I error rate), (c) the ability to accurately estimate the

[2] Analyses using simulated data indicated that model selection using the BIC without weighting tau by omega led to a slightly liberal selection criterion as manifested by a higher type I error rate and power (i.e., hit rate) than the OLS-BIC model selection. Note that our use of omega in equation 16 is intended to correct for overdispersion in the estimate of r_t'. In addition to using the form in equation 17, we also implemented a version in which omega = (N-k-1) / N and one in which r_τ was adjusted using the formula for adjusted r-square as given by 1-(1- r^2_τ)(N-1)/(N-k-1). As all of these variations led to approximately the same outcome, we choose to use (N-P-1)/N because it corrects for chance at the level of the experiment, rather than at the level of model selection. We note that a variety of alternative solutions to overdispersion based on traditional least-squares ML approach may also be appropriate corrections for overdispersion in GeMM, including variations of the BIC and the AIC that adjust for variance inflation factor, such as the QAIC (Burnham & Anderson, 1998) and the QBIC (Lebreton, Burnham, Clobert, & Anderson, 1992), though because these methods involve reliance on least-squares methods to compute a variance inflation factor their use seems inappropriate within the GeMM framework.

true population parameter values, and (d) the accuracy of predicting new observations (i.e., cross-validation).

Using these four criteria with simulated data, we evaluated three statistical modeling approaches:

1) *GeMM with model selection using the* BIC_τ' *(GeMM).* Our implementation of GeMM utilized the above equations and BIC_τ' to determine model fit. Parameter search was conducted using a genetic algorithm. The initial population of weight vectors for the genetic algorithm included 2^p weight vectors estimated via Least Squares estimation, where p is the number of predictor variables in the data set. In addition, 2000 vectors were generated by randomly perturbing the least-squared estimated weights and 4000 randomly generated weight vectors were also included in the initial population.[3] The GA was run for 10 generations. At each generation the 1000 vectors with the best fitness value (as determined by the BIC_τ') were selected for reproduction, with each new generation constrained to 4000 new members. In addition, the best 500 models from each generation were copied directly to the subsequent generation. The cross-over rate for reproduction was set to .85 and the mutation rate was set at .02.
2) *Least-Squares with model selection using the Bayesian Information Criterion (OLS-BIC).* Our implementation of OLS-BIC involved utilizing the genetic algorithm for parameter search, and using the BIC as the fitness function. Our implementation of the BIC used $1 - R^2$:

[3] The inclusion of the random weight vectors helps prevent the GeMM from settling into a local minimum dictated by the OLS solution. Although the inclusion of the least-squares regression weights improved the performance of the OLS-BIC model, these least-square weights had little impact on the performance of GeMM. Indeed, we implemented two different genetic algorithms for searching the parameter space, and both yielded equivalent patterns.

BIC = Nlog(1-R^2)+klog(N), where R^2 is the squared multiple correlation, N is the number of observations, and k is the number of parameters in the model.

3) *Least Squares regression with model selection using the Wald statistic (OLS-Wald).* Our implementation of OLS-Wald involved selecting parameters to include in the model based on significance testing using a Wald test with α=0.05.

The above three approaches were evaluated on data generated from a linear equation with a continuous multivariate distribution and six predictors, three of which were null, and where the other three had coefficients of .5, .3, and .2. Thus, the base equation generating the data for the linear environment was,

$$Y = .5X_1 + .3X_2 + .2X_3 + 0X_4 + 0X_5 + 0X_6 + 1 + e \qquad (19)$$

with $e \sim N(0, 1)$. For each simulation run two samples were created: one sample was used to estimate the model parameters, and the second was used for cross-validation. We manipulated sample size across 3 levels (N = 50, 100, and 250). These analyses were then repeated using a non-linear environment, which was created by using equation 19 but then raising Y to the power of 5 (Y^5). All analyses included 200 simulation runs. The GeMM and OLS-BIC models were conducted with MatLab code available from the authors; the OLS-Wald analyses utilized the built-in MatLab multiple-regression function.

Model comparison using simulated data

Recovering latent data structures. Figure 1 plots the performance of GeMM in terms of the hit and false-discovery rates compared to both versions of least-squares regression for the linear environment. Two observations should be evident. First, when data conform to strict linearity and multivariate normality, both versions of LS regression have slightly better statistical power than GeMM. Thus, as should be the case, GeMM is inferior to the linear model when the

data are actually linear, though these advantages are rather modest. For example, when $N = 100$, GeMM shows equivalent power for detecting the parameter with the strongest relationship ($\beta = .5$), and nearly equivalent power for the second strongest parameter ($\beta = .3$). This is remarkable given that GeMM ignores the metric properties of the data and fits only information contained in the ranks.

Next we investigated GeMM in comparison to OLS under the realistic assumption that data do not typically conform to linearity. Figure 2 plots the results of the simulations for the non-linear environment. In this case, GeMM shows a substantial advantage over the linear model across all sample sizes. For example, for N = 100, GeMM shows an improvement of 24% over both versions of OLS in detecting the second strongest parameter ($\beta = 0.3$), and an improvement of 24% over OLS-BIC, and a 15% improvement over OLS-Wald, for detecting the weakest parameter ($\beta = 0.2$). Thus, the assumption of linearity made by multiple regression, which is advantageous for the linear environment, is a liability for modeling non-linear environments. Importantly, inasmuch as the nature of data in psychological science is non-linear, the assumption of linearity will be a liability more often than not. GeMM, on the other hand, is robust to non-linearity, while showing false discovery rates similar to OLS-BIC.

Can GeMM provide relatively accurate estimates of the metric population parameters?

Researchers are often concerned with how *much* a particular variable contributes to the overall predictive accuracy of the model. Therefore, it is useful to know whether GeMM can provide reasonable estimates of the population parameters. In other words, how accurately can GeMM recover the metric parameters that created the observed data?

Figure 3 plots the mean normalized β's estimated by each model.[4] The population parameters correspond to .5, .3, and .2 for the first three parameters, and zero for the last three parameters. GeMM's estimates approximate the population parameters. Moreover, GeMM's estimates of these parameters are robust across the environments.

How can GeMM recover the population parameters without capitalizing on the metric properties of the data? The answer to this question stems from the fact that the number of constraints on the rank order increase exponentially with sample size – a property of the method of paired comparisons that Shepard (1962; 1966) argued allowed ordinal multidimensional scaling to approximate metric properties of the data. With reasonable sample sizes, the number of constraints on the ordered relationship is quite high. For example, with N=100 and no ties, there are 100(100-1)/2 = 4,950 paired comparisons required for determining the best parameters for optimizing the rank order correspondence between Y and $\hat{Y}$. As the number of observations in the rank order increases, the less freedom each value has to vary about its true value. Thus, as N increases, more precise estimates of the β's are required for minimizing rank order inversions. In the limit, these estimates should converge toward the true population parameters. Note, however, that the β's required for minimizing rank order inversions need not correspond exactly to the β's required for minimizing squared error. Therefore, GeMM's estimates are not guarunteed to be unbiased estimates of the population parameters, though inspection of Figure 3 would suggest that they are close approximations.

Predictive accuracy.

While Figures 1 and 2 clearly show that GeMM shows equal or better statistical power than least-squares regression techniques, how does it do in predicting new observations? To

[4] The parameter value was set to zero for any parameter not recovered by the model.

assess this, we compared GeMM to OLS-BIC and OLS-Wald using cross validation. The results of these comparisons are presented in Figures 4 and 5, which plots the proportion of concordances out of all pairwise comparisons, p(c)=C/[N(N-1)/2]), and the estimated tau for a sample size of 100. As can be seen, GeMM shows slightly less predictive accuracy in the cross-validation sample compared to both OLS procedures in the linear environment, but showed modest improvement in predictive accuracy in the non-linear environment. The principle reason that GeMM shows better predictive accuracy in the non-linear case is because it is better at identifying the true model underlying the data. Taken together, these results indicate that GeMM shows little loss of predictive accuracy in the linear environment, despite the fact that it is designed to model only monotonicity, and even shows better predictive accuracy than the linear model for non-linear environments.

A natural response to our demonstration above is that the linear model does poorly in the non-linear environment because the appropriate model is indeed non-linear. We of course agree with this position, but add that it is not always clear when data or the corresponding underlying phenomena are nonlinear. Moreover, even if one detects non-linearities within his or her data, it is virtually impossible to know precisely what the form of the non-linearities are and to verify that the non-linearities are representative of the larger population of scores , as opposed to being idiosyncratic (ie., random) properties of ones dataset. Just as few real datasets truly reflect multivariate normality (Micceri, 1989), it is also likely that few real datasets truly reflect linearity, and even modest, and perhaps even non-obvious departures from linearity can undermine statistical conclusions.

To illustrate this later point we reanalyzed data from Engle et al. (1999). Engle and his colleagues were interested in evaluating the factor structure of various measures of cognitive

ability. Our use of Engle et al's data is somewhat different: We were interested in using their data to evaluate whether GeMM had any practical advantages over various LS alternatives when applied to data where the true underlying statistical relationships is unknown.

Engle et al. (1999) had 133 participant's complete 12 measures of cognitive ability. To reduce computational time, we used only 10 of the 12 available variables. These included two measures of fluid intelligence (Ravens Progressive Matrices, Cattell Culture Free test), two standardized achievement tests (verbal SAT and quantitative SAT), three measures of short-term memory span (forward span, backward span, keeping track), and three measures of WM-span (operation span, reading span, and counting span). The use of cognitive ability data for evaluating GeMM was motivated by the assumption that, amongst real-world measured variables, measures of cognitive ability generally show reasonable distributional properties and approximately satisfy the assumptions of multivariate normality.

As a first step, we analyzed the data to determine whether it met multivariate normality. The results of these analyses were somewhat ambiguous: the Henze-Zirkler test for multivariate normality ($p=.12$) and the Mardia multivariate test of kurtosis ($p=.09$) failed to reveal significant departures from multivariate normality, whereas the Mardia test for multivariate skewness was significant ($p<.05$).[5] Thus, we concluded that whatever departures from multivariate normality were present in the data they were modest at worst, and did not justify transformation.

We examined the statistical power of GeMM versus the two LS approaches used on the simulated data, plus an additional four approaches (see Appendix A), including two versions or robust regression, Bayesian regression with normal priors, and ridge regression.

[5] Matlab code for the Henze-Zirkler and Mardia tests was obtained from the matlab file exchange (Henze-Zirkler test, Trujillo-Ortiz, Hernandez-Walls, Barba-Rojo & Cupul-Magana, 2007; Mardia test, Trujillo-Ortiz, & Hernandez-Walls, 2003).

We randomly sampled 50% of the total sample to form an estimation sample, and used the remaining 50% as the hold out, or cross-validation sample. For each run, we estimated the model on the estimation sample then applied the statistical model to predict observations in the cross validation sample. This was repeated 200 times for each approach. For the sake of comparison, we used the model based on the full sample as the criterion for evaluating statistical power. Using the full sample, all approaches identified a two-parameter model consisting of Q-SAT and Cattell's culture fair test as the best-fit model for predicting Raven's Progressive Matrices. However, how did the three techniques compare when the estimation sample was reduced by 50%?

The results comparing GeMM to the six alternatives based on least-squares are presented in Table 1, which provides the probability of recovering each parameter, and Table 2, which provides the fit and cross-validation statistics. As should be clear from examining Table 1, GeMM shows a substantial power advantage over OLS-BIC, but especially over OLS-Wald. The robust, Bayesian, and ridge procedures fared better than OLS-Wald, but also under-performed relative to GeMM and OLS-BIC. In fact, GeMM showed roughly a 16% increase in power over its nearest competitor (OLS-BIC) for detecting quantitative SAT, though a slight decrease in power (by roughly 3%) for detecting the Cattell culture fair test. Interestingly, closer inspection of these two variables revealed that quantitative SAT was slightly non-linearly related to Ravens (the best fit quadratic function accounted for nearly 5% more variance than the linear function when Q-SAT was regressed onto Raven's), whereas Cattell showed little non-linearity.

Turning now to predictive accuracy, how well did the various procedures fare in cross validation? These results are presented in Table 2. As can be seen, GeMM outperformed its competitors in terms of monotonic prediction, tau and p(c), while yielding a value for the

multiple r that approximated the best-fit linear model (OLS-BIC) and which was higher than the multiple r obtained by OLS-Wald, the robust procedures, and Bayes. Thus, despite the fact that GeMM ignores information captured by the metric fit, it predicts as well or better than the least-squared alternatives. The only procedure to outperform GeMM was ridge regression, which is a more complex modeling approach that uses all of the available predictors.

What have we learned from the above analyses? First, it is clear that GeMM seems to be a useful tool for statistical modeling. Obviously, under conditions in which the data is actually linear, LS regression procedures will be more powerful. However, under the more realistic conditions in which the data is monotone, but non-linear, GeMM is more powerful. Second, GeMM is robust to departures from linearity in a way that standard regression procedures are not. Thus, there is no need to test out various transformations on the data to see which one "works" and no need for outlier deletion. The implications of these first two points, however, go well beyond simple lessons for statistical modeling, as they go to the heart of the fundamental goal of science: Making assumptions about the nature of data that do not actually hold, just so a statistical analysis can be performed, can ultimately affect ones theoretical description of the corresponding phenomenon. The evidence taken in support of a theory about phenomena should not rest on questionable assumptions required for the statistical analysis, nor should such evidence be conditional on a particular transformation. Ideally, statements about phenomena should require as few assumptions as necessary. GeMM allows one to model his or her data without recourse to data transformation or the default assumption of linearity. A final finding from the initial set of analyses above is that GeMM is as good, or better, in predictive accuracy than linear regression – a finding that is of particular relevance for evaluating models of behavioral decision making.

GeMM as a model of choice.

Having developed GeMM as a tool for describing and analyzing data, we now turn to explicating GeMM within the context of theoretical models of decision-making. Of note, we make no claims as to the psychological plausibility of GeMM as a *mechanistic-level* description of decision behavior. Rather, we view GeMM as serving two complementary purposes. One purpose concerns prescriptive decision making and the other the description of decision making at the computational level. When used for prescription GeMM can be viewed as a model of choice that combines information in a way that respects the monotonic properties of the data when allotted unlimited time and computational resources. The second purpose concerns descriptive modeling and judgment analysis; GeMM allows one to describe cue utilization, in the same way that multiple regression has been used within the context of the Lens model (Hammond et al., 1964; Karelaia & Hogarth, 2008). The use of GeMM in this way has allowed us to identify important theoretical issues that hitherto have been largely neglected in evaluating heuristic models of human judgment. Before addressing these issues, however, we first describe a few alternative approaches to modeling decision making behavior.

Egon Brunswik's (1951; 1955) Lens model framework is one of the most widespread approaches to modeling judgment and decision making. One of the primary goals of the framework is to understand how people utilize information in inference tasks. Hammond et al. (1964) were among the first to use multiple-linear regression to parameterize and estimate the components of the Lens model, and the use of the linear model for estimating parameters in the Lens model subsequently became commonly accepted practice (Hammond, et al., 1964; Hartwig, & Bond, 2011; Hastie, & Kameda, 2005; Hursch, Hammond, & Hursch, 1964; Steinmann, &

Doherty, 1972; Tucker, 1964; York, Doherty, & Kamouri, 1987; for a review see Karelaia, & Hogarth, 2008).

The Lens model makes two key assumptions regarding decision behavior. First, it is assumed that the decision maker has some knowledge of the relationship between the criterion of interest (Y) and the cues (x_i's) as they exist in the environment. Second, judgment about the criterion variable is assumed to be based on an additive combination of some or all of the cues known to the decision maker. Given these first two assumptions, it is easy to see how the linear model can be used to parameterize or estimate which cues a participant utilized in making a set of judgments, and how much weight each cue was given. The configuration of cues and their respective weightings are taken as the decision maker's *policy*. This policy can then be compared to the optimal policy, as given by the parameters estimated on the environment. Clearly, the use of the linear model to characterize a decision maker's policy entails all of the assumptions involved in multiple-linear regression, including that of linearity. Although in the General Discussion we describe the Lens model components in detail and reconceptualize them within the GeMM framework, we do not use the Lens model in its entirety here. Instead we evaluate its inference engine (i.e., the linear model) in comparison to GeMM on their ability to cross-validate— to predict a hold-out sample with cue weights estimated on an estimation sample.

As an alternative model of decision making**,** Gigerenzer et al. (1991) proposed the Probabilistic Mental Models (PMM) theory for modeling the environment-behavior interaction. Although this framework builds on the core ideas of Brunswik's (1955) Lens model, it does not rely on multiple linear regression. Instead, PMM postulates that people possess a toolbox of heuristic mechanisms that exploit the properties of the environment. One of these heuristic mechanisms is Take the best (TTB), which has received a great deal of attention in the literature

(see Gigerenzer & Goldstein, 1996; for a critique, see Dougherty, Franco-Watkins, & Thomas, 2008). Take-The-Best is a three-step algorithm that involves (1) ordering predictor variables according to their predictive accuracy (defined as cue validity, see below), (2) searching the predictor variables from most to least predictive, and (3) terminating search when a cue is found that discriminates between alternatives in a paired comparison.

Specifically, TTB defines the 'best' predictor in terms of cue validity, where cue validity is given by: $v_i = p[t(\mathrm{a}) > t(b) \mid (c_i(a) = +) \text{ and } (c_i(b) = -)]$. The validity ($v_i$) of cue i, denoted c_i, on target variables a and b, t(a) and t(b), is given by the relative frequency with which $t(a) > t(b)$ given that a is positive on cue i and b is negative in reference class R. The terms $t(a)$ and t(b) correspond to the value of the target variable for case a and b, respectively, where the choice set, {a,b}, is a paired comparison. Note that the formula for cue validity can be re-written in terms of the number of paired comparisons in which the predictor and criterion are in concordance divided by the number of paired comparisons that are either concordances (C) or disconcordances (D): v=C/(C+D)[6]. It should be clear that the formula for cue validity is an index of monotonic fit, in the same spirit as tau, but without penalizing the numerator for disconcordances and ignoring ties in the denominator. Thus, TTB can be viewed as an alternative model to GeMM for capturing monotonicities.

Take-the-best proceeds by evaluating the cue values in the order of their validity. If the cue discriminates then TTB chooses whichever choice alternative has the positive cue value. If the cue does not discriminate then TTB evaluates the next most valid cue, and so forth. If none of the cues discriminate, then an alternative is chosen at random.

[6] Cue validity re-written in terms of concordances and disconcordances, C/C+D, is similar to the well-known gamma correlation, in which the rank order correlation between two variables is defined as (C-D)/(C+D)(Goodman & Kruskal, 1979).

TTB differs from the linear model in two respects. First, TTB uses only a single cue at a time whereas the linear model integrates across cues. Second, the fit criterion for TTB is based on monotonicity, whereas the fit criterion for the linear model is based on least-squares. Thus, any differences in fit when comparing the linear model to TTB could in principle be due to whether the cues were integrated or not, or due to differences in the fit criteria, or both. Indeed, our analyses of GeMM in the first half of this paper already showed that the use of the linear model in predicting observations in a non-linear monotone environment leads to sub-optimal choice behavior. Of course, it remains to be seen whether GeMM outperforms other algorithms that model monotonicity, such as TTB.

We evaluated the predictive accuracy of GeMM relative to the linear model and TTB within the context of addressing three theoretical issues. The first issue centers on the topic of parsimony. An important construct in evaluating descriptive models of human choice involves the need for such descriptions to respect the cognitive limitations of the human decision maker. Simon (1956) dubbed this bounded rationality, and it has become an increasingly important consideration in the decision-making literature as the focus has turned to describing decision making in terms of heuristic mechanisms. We illustrate that modeling the fit/parsimony tradeoff with variable selection criteria such as the BIC can shed new light on the issue of bounded rationality and its relation to models of cue integration such as the linear model and GeMM.

The second issue addresses the need to minimize incorrect choices while maximizing correct choices. We illustrate that models that focus solely on maximizing correct choices, such as TTB, can yield a seemingly paradoxical effect: The number of incorrect choices can increase at a disproportionately greater rate than correct choices. The use of BICt' for model selection can prevent this from happening.

The third issue concerns the very simple proposition that choice models based on a linear least-squares fit criterion can be ill equipped for accurately modeling choice behavior. Note that GeMM's goal of maximizing monotonic fit (ie tau) is consistent with the goal of maximizing choice, but that the parsimony correction ensures that there is a trade-off between choice accuracy and model complexity. Thus, the use of the BIC may come at a cost of choice accuracy, even though the goal of GeMM is to maximize tau. The contrast with the linear model is that linear least-squares regression does not directly model choice behavior, but rather infers choice accuracy from the least-squares fit. We illustrate that the best-fit linear model can yield choice accuracy far below that of GeMM, even in cases in which the linear model accounts for a substantially greater proportion of the variance.

Issue 1: A rational model of behavioral choice should trade fit for parsimony.

A parsimonious algorithm is one that uses as little information as necessary to do the job. Simplicity is an important element of Simon's bounded rationality and one that we model. Ideally, one wishes to maximize the fit or predictive accuracy of an algorithm, while simultaneously retaining parsimony. The notion of parsimony has been a central construct of the fast and frugal heuristic tradition. Specifically, Gigerenzer and Goldstein (1996) and others argue that fast and frugal heuristics require less information upon which to make a decision. Within this tradition, parsimony (i.e. frugality) is defined locally for each paired comparison – that is, the number cues one must search in order to discriminate between two objects. We refer to this form of frugality as *local parsimony* because it applies locally to individual pairs of objects.

Yet, local parsimony (or frugality) as used in the study of fast and frugal heuristics is quite distinct from the notion of parsimony as used in the statistical and computational modeling literatures. Within these literatures, parsimony is defined over the entire class of objects in the

judgment set – that is, how many cues are needed to maximize the fit to all of the data (i.e, all paired comparisons). This sort of parsimony is made explicit in model selection procedures such as the BIC, AIC, and RIC, and is indexed by k in the BIC. We refer to this form of parsimony as *global parsimony*, because it applies globally across the entire data set.

Unfortunately, past evaluations of TTB and other fast and frugal heuristics have largely neglected the issue of global parsimony. While the definitions of local and global parsimony seem to imply different metrics for assessing the parsimony of a model, they are closely related. For example, within the context of GeMM and multiple-regression, the local parsimony of the model is always identical to its global parsimony. That is, for each and every paired comparison, GeMM and OLS use an additive combination of k cues. Thus, both the local and global parsimony of GeMM and OLS is equal to k. Within the context of TTB, the global parsimony of the model sets the upper bound for local parsimony. That is, TTB may require k cues to make inferences across the entire class of objects (i.e., all possible paired comparisons), but may only use k_i* cues on the ith paired comparison, where k_i* <= k. Global parsimony, therefore, is given by max(k_i*), or the maximum number of cues required for any single paired comparison. In most analyses of TTB, researchers report the mean value of k*, the average local parsimony, which will often be less than k, the global parsimony. The question we address here is: how does TTB perform relative to GeMM and OLS when global parsimony (k) is held constant?

To illustrate the importance of modeling global parsimony, we reanalyzed the dataset used in the initial presentation of the take-the-best heuristic – the city's data set. This dataset consisted of all cities in Germany with more than 100,000 inhabitants (i.e., the 83 largest German cities by population), their population value (the criterion), and 9 binary predictor cues (Gigerenzer & Goldstein, 1996). The predictor cues included: 1) National capital (Is the city the national

capital?), 2) Exposition site (Was the city once an exposition site?), 3) Soccer Team (Does the city have a team in the major league?), 4) Intercity train (Is the city on the Intercity line?), 5) State capital (Is the city a state capital?), 6) License plate (Is the abbreviation only one letter long?), 7) University (Is the city home to a university?), 8) Industrial belt (Is the city in the industrial belt?), 9) East Germany (Was the city formerly in East Germany?).

We used a split-half cross-validation procedure to examine the performance of the models using three metrics: The value of BICt', Kendall's tau, and the proportion of concordances. The BICt' for GeMM was calculated as shown above, where k is based on the number of parameters included in the model. The BICt' for TTB was computed by calculating the equivalent of Kendall's tau from the concordances, disconcordances and ties derived from the TTB algorithm. The value of k used in the BICt' formula was determined by the total number of cues needed to resolve all pairwise comparisons. In this sense, k is actually identical to the max k_i*. For the analyses presented below, we varied the value of k parametrically from 1 to P, where P was the number of predictors in the dataset. For example, when k = 1, each model was forced to choose the best one parameter model, when k = 2, each model was constrained to find the best 2 parameter model, and so forth. For each value of k, we conducted 100 monte carlo runs, with each run consisting of estimation and cross validation samples. The best fitting parameters for GeMM and OLS were estimated on the estimation sample using the genetic algorithm (restricting the algorithm to including only k non-zero parameters) and then applying this model to the cross-validation sample. For TTB, the estimation sample was used to estimate the cue validity of all P predictors, and then to select and order the k best 1-parameter models according to cue validity. The cue ordering of the selected k 1-parameter models was then applied to the cross validation sample.

Of note, within TTB one cannot directly estimate a Pearson's r. However, given that TTB yields all of the components necessary to compute Kendall's tau, it is possible to use the tau to r conversion as a proxy for the Pearson's r, which allows us to use equation 18 for BICt', making a straight-forward comparison between TTB, GeMM, and OLS.

Model comparison results. Figure 6 plots the mean values for BICt' as a function of the number of predictors included in the model. Two things are of note. First, as illustrated by the U-shape curve there is a clear trade-off between number of cues and the value of BICt'. According to all three algorithms, on average the best model (lowest BICt') derived from the estimation sample included 2 parameters. More generally, across the three algorithms, only models with 4 or fewer parameters consistently showed negative values for BICt'. Thus, even though accuracy continues to improve as more cues are utilized, model selection using the BICt' leads us to prefer models with 4 or fewer parameters, with the best overall fit shown for models with 2 parameters. Second, GeMM provides the best overall fit to the estimation sample, as evidenced by lower values of the BICt' (more negative).

Turning now to the assessment of model accuracy, Figure 7 plots the mean percentage of concordances, whereas Figure 8 plots mean values of tau as a function of number of cues for both the estimation sample (top panel) and the cross-validation sample (bottom panel). Consistent with the analysis of BICt', GeMM outperforms TTB and OLS-BIC on the estimation sample on both measures for all values of k. The results for the cross-validation sample illustrate two interesting findings. First, GeMM clearly outperforms both OLS-BIC and TTB for models with fewer parameters. In fact, the predictive accuracy of GeMM (percentage of concordances and tau) exceeds that of TTB and linear regression when the algorithms are restricted to $k < 4$, indicating that when disconcordances are factored into the assessment of model fit, GeMM is the

clear winner. Note that GeMM is the only model optimized on tau, so it is perhaps unsurprising that it outperforms TTB and OLS on prediction. However, the critical point here is that within least-squares regression, one assumes interval level data. Under this assumption, one would expect that maximizing fit at the interval level, presupposes that the ordered relationship are well accounted for. As we show below, however, this isn't necessarily the case. Moreover, GeMM still does better than both TTB and OLS in terms of percent correct. Thus, after equating the three models for number of parameters and trading fit for complexity, GeMM appears to show a reasonable advantage in predictive accuracy. Interestingly, the advantage for GeMM is not as evident when evaluating tau (Figure 8).

Second, when allowed to search all 9 cues TTB outperforms both GeMM and OLS-BIC in terms of the proportion of concordances. However, the Kendall's tau for the three parameter GeMM model (M=0.463) is actually nominally higher than tau for the nine parameter TTB model (M=0.461), suggesting that a three-parameter GeMM model is nearly good as a 9 parameter TTB model when evaluated in terms of the monotone correlation. Why does TTB outperform GeMM and OLS on percent concordance but not tau? This is the focus of Issue 2.

Issue 2: A rational model of behavioral choice should maximize correction decisions, but minimize incorrect decisions.

Careful inspection of the data in Figures 7 and 8 (cities data) highlights a curious finding: The percent of correct inferences continues to increase even after the value of tau plateaus. How is this possible? The answer lies in the nature of disconcordances, or *incorrect* choices. Consider the following two equations:

$$P(C) = C/(C+D+T_P+T_C) \quad (20)$$

$$Tau=(C-D)/(C+D+T_P+T_C) \quad (21)$$

Where C, D, and T are defined as concordances, disconcordances, and ties, with ties on the predictor denoted T_P and ties on the criterion denoted T_C. C corresponds to a correct choice, whereas D corresponds to an incorrect choice. Equation 20 gives the percent of correct choices, as indexed by the concordances. Equation 21 is a variant of the version of Kendall's tau in equation 9. For any class of objects, there are N(N-1)/2 paired comparisons. Any paired comparison that is not a C or a D is either a T_P or a T_C. As should be clear, the top formula includes no penalty for an incorrect choice, whereas the bottom does. Thus, whether a cue leads to a tie or to an incorrect choice does not affect the value of P(C). To be clear, cues that discriminate between any pair of objects can produce a C or a D; for each paired comparison that is resolved as either a C or a D, the value of T_P is decreased by one. Thus, theoretically, it is possible for P(C) to increase monotonically even if the preponderance of paired comparisons resolved by a cue are disconcordant. We argue that any rational decision algorithm should be penalized for making incorrect decisions. P(C) therefore provides one with a one-sided assessment of predictive accuracy, and can only be interpreted in the context of a metric that takes into consideration incorrect choices, such as tau. In this section, we illustrate that the TTB heuristic ignores the impact of incorrect choices, whereas GeMM does not when the value of BICt' is minimized. As we illustrate below, following a decision strategy that focuses solely on maximizing P(C) can come at a cost of making a disproportionate number of incorrect choices. However, explicitly accounting for incorrect choices in model selection using BICt' can remedy this problem in both GeMM and TTB.

To illustrate, consider an environment in which one wishes to predict Y from two variables, A and B. Suppose further that cue A results in 100 concordances, 25 disconcordances, and leaves 200 dyads tied on the predictor. So, with cue A, PC = 100/(100+25+200)=.307, and

Tau = (100-25)/ (100+25+200)=.231. Now consider what happens when a second variable is added to the prediction equation that results in 20 additional concordances, but 25 additional disconcordances: PC increases to 0.369 (120/(125+50+155), but tau decreases to .215 ([120-50]/[125+50+155]). In other words, the 6% increase in correct choices was accompanied by a nearly 8% increase in incorrect choices. Costs and benefits held equal, it makes little sense for a rational decision maker to use cue B because it will result in more incorrect choices than correct choices – in this case, ignorance is bliss. Yet, under TTB, cue B would be used and the disproportionate increase in incorrect choices would be tolerated. In contrast, when using BICt' for model selection within GeMM, any cue added to the model must result in a greater number of concordances relative to disconcordances for it to be justifiably included in the model.

As a real-world demonstration of the example given above, we re-analyzed a dataset previously used to evaluate the accuracy of TTB that involved predicting professor salaries from a set of 5 cues, including rank, years in current rank, year degree was earned, sex, and highest degree obtained (see Gigerenzer et al., 1999; Lee & Cummins, 2004). To allow TTB to use all 5 cues, the data were dichotomized using a median split. Using the full sample (N=52), we ran GeMM, OLS-BIC, and TTB constraining the algorithms to using k cues, where k was varied from 1 to 5. The results of these analyses are presented in Figures 9 and 10. Two observations should be evident. First, there is an inverted u-shaped function relating tau to the number of parameters for GeMM, OLS-BIC and TTB, while there is a monotonically increasing function relating proportion of concordances to the number of parameters. This suggests that the inclusion of more parameters in any of the models does not automatically lead to an improvement in choice behavior when incorrect choices are considered. In fact, allowing TTB to search all 5 cues instead of the 4 best cues led to an additional 31 correct choices (C's), but at a cost of 56

incorrect choices (D's). Second, and perhaps more interesting, the Pearson's r generally showed a monotonic increase as a function of the number of parameters. Thus, if one were to simply maximize the number of concordances or the Pearson r, it would result in a disproportionately large increase in incorrect choices. In contrast, model selection using the BICt' obviates this problem – the best-fit model in GeMM is the one that maximizes the trade-off between correct and incorrect (pair-wise) choices.

Issue 3: Decision models based on linear least-squares optimization can yield relatively poor choice behavior.

As suggested earlier, the parameters required for minimizing the squared error loss function need not correspond to the parameters needed for minimizing rank order inversions. That is, it is possible for the best-fit linear model to account for the most variance, in the traditional sense, and yet do relatively poorly at recovering the rank order solution. Certainly, one should question the validity of any metric solution that fails to recover the more primitive property of ordinality, even if it shows a large R^2. This point has important implications for the standard applications of Brunswik's Lens model that minimizes the squared error loss function.

To illustrate this point, we re-analyzed the Cities dataset using both GeMM and OLS-BIC. Using the full dataset, the best-fit model based on OLS-BIC consisted of three parameters, including the Soccer Team, Exposition site, and National Capital cues, and accounted for 84.8% of the variance, but yielded a Kendall tau of 0.517, and concordance rate of only 38.2%. The best-fit GeMM model also consisted of three parameters, and included the Soccer team, Intercity train line, and Exposition site cues. In contrast to the best fit-linear OLS-BIC model, the best fit GeMM model only accounted for 24.2% of the variance. However, inspection of the P(C) and tau correlation revealed that GeMM yielded a much higher rank-order correspondence than

OLS, with tau = 0.548 and a concordance rate of P(C) = 55.4%. Thus, despite the fact that OLS-BIC accounted for over 3 times more variance than GeMM, it performed substantially worse in predicting the rank orders and lead to nearly 17% (55.4%-38.2%) *fewer* correct choices. The striking contrast between GeMM and OLS-BIC demonstrates a simple point: The least-squares fitting criterion is not designed with the goal of choice maximization. Perhaps even more illuminating is the fact that the best-fit OLS-BIC model was actually qualitatively different than the best-fit GeMM model, in that they consisted of different predictors!

While the above findings are of interest for purely statistical reasons, they are also of central importance for the development of prescriptive and descriptive models of choice. The linear model has long been used as a paramorphic model of human decision making and as the normative standard against which human accuracy is compared (Hogarth, & Karelaia, 2007 Karelaia & Hogarth, 2008). However, it is clearly the case that a decision agent operating under the principle of maximizing choice accuracy may utilize different information than a decision agent operating under the principle of linear least-squares.

General Discussion

We believe that the work presented in this paper illustrates the potential benefits of bringing statistical models in line with the precision of psychological theory and the (non-linear) nature of the manifest relations in much behavioral data. Using both simulated and real data we demonstrated that GeMM successfully models non-linear relations without the need to make (overly) precise assumptions about functional form. Moreover, when the functional relation in the data is strictly linear, GeMM suffers negligible losses in terms of power and predictive accuracy. Thus, there seems to be much to gain and little to lose by adopting GeMM as a model of statistical inference. Beyond the use of GeMM as a statistical model, we argued that GeMM

could be used as a prescriptive and descriptive framework of judgment and decision making, and then illustrated this usefulness in indentifying several critical issues within behavioral decision theory.

Given the potential for linearity's and non-linearity's in one's data, it is instructive to compare the costs and benefits of assuming linearity with the costs and benefits of modeling data under the less-restrictive assumption of monotonicity. That is, what does one give-up by adopting the less-restrictive assumption of monotonicity when the true state of the data is linear? And, what does one give-up by adopting the assumption of linearity when the true state of the data is non-linear? The analyses presented in this paper suggest that the answer to the first question is 'very little'. As we showed in the first set of analyses, even when data meet the assumptions of linearity, one has little to lose by adopting the less-restrictive assumption of monotonicity and modeling data with GeMM. In contrast, the cost of assuming linearity when linearity does not hold can be considerable. Indeed, as we showed using both simulated and real data, applying the linear model to non-linear data can lead to a substantial reduction in statistical power and poorer predictive accuracy.

One key insight gleaned from our development of GeMM is that it is possible to model any monotonic statistical relationship without making a priori assumptions of functional form. This is made possible by exploiting the transformation on Kendall's tau within a model-based framework. The tau to r transformation essentially allows one to model any monotonic nonlinear relationship *as though* it was linear, but without fishing for the transformation that best approximates linearity. This is a major insight, in that it frees the researcher of deciding whether or which transformation to use on the criterion, and eliminates the need to conditionalize statistical conclusion on the choice of a particular transformation, for example when two

different transformations yield different conclusions. Moreover, because GeMM operates at the level of rank correlation rather than squared deviations, the presence of outliers or extreme scores is of less concern. In our view, using data transformations to reshape the data so that they appear to fit the assumptions of a statistical model is a bit like forcing a round peg into a square hole. Rather than changing the data to fit the statistical algorithm, it makes more sense to use an algorithm appropriate for the data at hand (see Cliff, 1996). GeMM is an appropriate approach to statistical modeling of psychological data because it does not require one to adopt overly restrictive assumptions about functional form that may or may not be represented in the data or population.

An important property of GeMM is that the fit criterion is based on the minimization of rank-order inversions. The idea of minimizing rank inversions is consistent with the goal of many applied researchers who seem less concerned about predicting precise quantitative values than predicting relative values (i.e., rank order, cf. Cliff, 1993; 1996). For example, in personnel selection contexts, it is advantageous to select individuals according to the rank order of applicants, without regard to the quantitative properties of the distribution of scores (Schmidt, 1995; see also Campion, Outtz, Zedeck, Schmidt, Kehoe, Murphy, & Guion, 2001). Moreover, even when quantitative values are available, they are often re-expressed in terms of their ordinal or rank properties, as exemplified by the use of percentile ranks when interpreting standardized test scores such as the WAIS or GRE, and the use of linear rank selection in personnel selection decisions (Schmidt, 1995).

Given that the goal of many applied domains is really to identify ordinal relationships, it makes sense to model this property directly, rather than infer it from a least-squares solution. Importantly, in many real world applications even a small difference in accurately predicting

rank orders can yield large economic or health outcomes. For example, consider health policies designed to identify individuals most at risk for developing breast cancer, or policies designed to identify individuals prone to obesity or drug abuse. The difference of 2% in predicting the rank orders can mean the difference between catching or missing critical high-risk cases. As our analysis of the cities data under Issue 3 illustrates, it is possible for the linear model to substantially underperform GeMM in predicting rank orders, even when the value of R-square is substantially higher than that estimated by GeMM. Thus, a high-value of R-square derived from application of the linear model and assuming interval-level data may give one the illusion that the ordinal properties of the data will follow. However, this need not be the case since the least-squared solution is designed to minimize squared error, even if it comes at a cost of *mis*-predicting the ordinal properties of the data. In as much as prediction of ordinal properties is an applied goal, then in our opinion it makes more sense to model the ordinal properties directly, rather than to infer it from the best-fit linear model.

Aside from its implications for statistical decision making, our development of GeMM also highlighted a number of issues relevant to theories of behavioral decision making. For example, we highlighted the need to control for global model complexity when comparing different decision models. In as much as parsimonious models are preferred, parsimony should be explicitly accounted for when assessing model fit. We illustrated that the predictive accuracy a popular heuristic model of decision-making, take-the-best (TTB), depends on its global complexity (how many cues it is allowed to search), and that TTB actually does worse than GeMM when global complexity is explicitly modeled. Our conclusion, therefore, is that non-compensatory decision heuristics like TTB can be advantageous, but only when the amount of information (cues) is unconstrained.

A second issue highlighted by our development of GeMM is the need for decision models to explicitly account for incorrect choices as well as correct choices. In past work on take-the-best, the focus of researchers has been on the proportion of correct choices made by TTB. However, the singular focus on correct choices obscures the curious property that more 'accurate' models may also commit more incorrect choices. We illustrated that it is possible for the incremental validity of particular cues to be negative, in the sense that using these cues can lead to more incorrect choices than correct choices.

Finally, a third issue highlighted by GeMM is that decision models based on linear least-squares optimization can appear to perform well when assessed in terms of R^2, but yield relatively poor choice behavior. This finding illustrates one of the primary limitations of the linear model as a model of choice – it is not designed with the goal of predicting the ordered relations and therefore, will not tend to maximize (pair-wise) choice accuracy. Rather, its goal is to minimize squared error, even if that comes at the cost of choice accuracy.

While the above findings illustrate the contribution of GeMM to understanding models of decision making, the implications of GeMM for behavioral decision theory goes well beyond highlighting these three issues. Indeed, as we illustrate below, GeMM allows for a fundamental reconceptualization of a variety of theoretical frameworks within the decision sciences. We outline of few of these extensions next.

Extensions to models of behavioral choice.

The analyses presented throughout this paper illustrate that GeMM is useful both as a model for statistical inference and as a descriptive model of behavioral choice. However, these analyses have only scratched the surface of its generality. In this section, we outline some natural extensions of GeMM and illustrate its relationship to various models of behavioral choice.

The majority of the work presented throughout the second half of this paper has focused on two seeming different approaches to judgment and decision making: the Lens model, as proposed by Brunswik, and a heuristic-based approach, as advocated most recently by Gigerenzer & Colleagues (Gigerenzer, Todd, & The ABC Research Group, 1999; Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002). Both of these approaches can be re-conceptualized within the context of GeMM, and doing so highlights a number of relationships between heuristic algorithms and the Lens model approach, while also allowing us to identify novel variants of the heuristics.

The Monotone Lens model.

One of the most widely used frameworks for modeling decision behavior is the Lens model (Hammond, et al., 1964; Hartwig, & Bond, 2011; Hastie, & Kameda, 2005; Hursch, et al. 1964; Steinmann, & Doherty, 1972; Tucker, 1964; York, et al., 1987; for a review see Karelaia, & Hogarth, 2008). The Lens model utilizes multiple linear regression to model both the statistical relationships in the environment and how people use information in inferring properties of the environment (i.e., judgment analysis; for a review see Karelaia and Hogarth , 2008). Indeed, in a recent review of Lens model work, Karelaia and Hogarth (2008) identified over 200 published and unpublished studies using the lens model. This framework is presented in Figure 11 and Table 2 presents interpretations of model constructs. However, we argue that the common use of linear regression in modeling human judgment is misplaced because there are many potential sources of non-linearity's either in the environment (i.e., the true relationships) or in the decision making process (Einhorn, 1970; Einhorn, Komorita, & Rosen, 1972; Kim, Yang, & Kim, 2008).

What are the implications if the state of nature is in fact not strictly linear, but conforms to the less special case of monotonicity? In such a case, the linear model is the wrong model and is therefore estimating the wrong weights on the both the ecology and policy dimensions of the Lens model. Thus, deviations in the weights estimated on the environment, and the weights estimated on the judge may arise either because the weighting scheme adopted by the judge is ill-calibrated with respect to the environment *or* because the weights identified by the linear model are estimations based on phantom properties of the environment--properties that may not exist! That is, perhaps the linear model is the wrong statistical model against which to compare and model human behavior. In as much as the manifest data is monotonic, but not linear, the weights estimated by multiple regression may poorly characterize the judgment policy adopted by the decision maker or the optimal policy suggested by the environment, or both. As illustrated under Issue 3, it is even possible for the best-fit linear model to be qualitatively different from the best-fit monotone model.

The obvious solution to this problem is to adopt a modeling approach that captures monotonic properties, but does not assume strict linearity. Thus, we propose GeMM as a paramorphic model of human judgment and decision making. Table 3 recasts the common Lens model decomposition in GeMM terms. For the monotone Lens model in Table 3 we have chosen to use the Pearson correlation coefficient estimated via Kendall's tau. Although we could have just of as easily have captured the Lens model components with other measures of monotone association, the tau estimated Pearson correlation can be interpreted as the standard Lens model with the linear estimate being derived under the less restrictive assumption of monotonicity.

It is also important to point out that the traditional use of the Lens model has essentially ignored the issue of model complexity. Our monotone Lens model allows for a direct

comparison between any model from which paired comparison data can be obtained, including the linear model relying on the BICt'. Thus, the fit-parsimony tradeoff is explicitly considered in terms of the BICt'.

In summary, there are many compelling reasons to prefer a monotone lens model. First, *linearity* is a special case of monotonicity and as illustrated by the simulation work, GeMM shows little loss in statistical power when data is linear. The monotonicity assumption also allows for more flexibility and robustness as illustrated in simulation by GeMM's substantial power advantage as data becomes increasingly nonlinear, but remains monotonic. Finally, when assuming linearity and using least-squares optimization, one implicitly endorses a fit metric that tolerates rank inversions in order to minimize squared deviations. This tradeoff is not defensible in our opinion because rank is a more fundamental measurement property than distance.

Heuristics as monotone decision algorithms.

While GeMM can be considered a monotone version of the lens model, it can also be viewed as a general framework for organizing both compensatory and heuristic models of decision making. Specifically, several decision models are closely related to or can be expressed as special cases of GeMM; moreover, our development of GeMM highlights several variations on existing heuristic models and suggests new models.

Table 4 presents several decision models that fit within the GeMM framework: The lens model, Franklins rule, Dawes' rule, Take the best, and Minimalist. For example, Take-the-best can be viewed as a variant of GeMM in which one searches through a set of one-parameter models from best to worst until one model discriminates. That is, define a regression model based on GeMM with P parameters (cues). TTB is related to GeMM in that there are P one-parameter models nested within GeMM's P parameter model. TTB differs from GeMM only in

the sense that the decision maker is assumed to order the P 1-parameter models by validity and search them one at a time. Both Dawes's (the equal weights additive rule, or ADD) and Franklin's rule (weighted additive rule, or WADD) can also be considered special cases of GeMM. Both assume that the decision maker chooses between alternatives by combining cues. In the case of Dawes's rule, the cues are combined additively (summed) with equal weight. In the case of Franklin's rule, the cues are weighted by importance before summing (multiple regression is Franklin's rule where the weights are optimized via least-squares estimation). Pairwise choice in both models assumes that the decision maker chooses the alternative with the highest sum. WADD and ADD have been investigated extensively as models of choice (Broder, 2002; Broder & Schiffer, 2003; Payne, Bettman, Johnson, 1992, 1993). Conceptualized within GeMM, these are merely saturated models that assume that participants combine all possible cues (with different weighting schemes) in choosing between alternatives. Note that none of these heuristics explicitly attempt to maximize choice accuracy.

Also represented in Table 4 is the Borda count, which is a voting rule for social preferences in which N choice options are ranked by individuals, where the highest ranked alternative is assigned a ranking of N, the second highest ranked alternative is assigned a value of N-1, and so forth (Borda, 1784). The winner of a Borda count is the alternative with the highest summed rank. The Borda count is easily expressed as a special case of GeMM, where the predictor variables are viewed as voters, and each voter provides a ranking of all N alternatives. Within GeMM, any dataset can be expressed as a Borda count by (a) rank transforming the predictors, (b) assigning unit weights to the predictors such that each 'voter' is counted equally, and (c) summing across all predictors. The alternative with the highest summed ranks is the Borda winner. A natural extension of the Borda count suggested by GeMM that is not

represented in Table 4 is a weighted Borda count, where each 'voter' is weighted by a validity coefficient derived from a test sample. A reasonable application of the weighted Borda count would involve any prediction task where judges with various levels of expertise or knowledge are predicting the outcome of a contest. Weights estimated from an estimation sample could be taken as an index of each judge's validity (c.f., Ho, Hull, & Srihari, 1994).

In addition to providing an organizing framework for the above choice rules, GeMM also highlights several novel variations of these heuristics. In particular, each of the models described above can be coupled with a model selection procedure that trades fit for parsimony. For example, TTB-Select and Minimalist-Select use the BICt' to reduce complexity. As well, these models can be naturally extended to include any possible parameterization, so that the TTB search algorithm can operate on any ordered set of models containing k>=1 parameters (TTB-GeMM and Minimalist-GeMM). For example, under a GeMM formulation, one can order all possible 2-parameter models and search them sequentially using TTB's search and stopping rules. Precisely how to define the value for number of parameters (K) and how to select amongst competing models is yet unclear, so these are left blank in the table. Nevertheless, these are valid decision heuristics that should be explored.

Summary

Fundamentally, the work presented in this paper illustrates the potential benefits of modeling psychological data without recourse to making strong, and often untenable, assumptions about ones data, and without the need for employing data intervention strategies aimed at bringing the *data* in line with the assumptions of a statistical model. Rather, we argue that a more appropriate approach to data analysis is to utilize a modeling approach that respect the fundamental properties of one's data. GeMM accomplishes just this. GeMM is based on the

principle of monotonicity, and as such requires much less restrictive assumptions about the modeled relationship. Thus, GeMM permits one to capture the (non-linear) nature of the manifest relations that exist in much behavioral data and to model the data at a level of precision that is consistent with the precision of most psychological theories.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, **2,** 396–408.

Anderson, N. H. (1968). A Simple Model for Information Integration. In R.P. Abelson, E. Aronson, W.J. McGuire, T.M. Newcomb, M.J. Rosenberg, & P.H. Tannenbaum (Eds.), Theories of Cognitive Consistency: A Sourcebook. Chicago: Rand McNally.

Anderson, N. H. (1970). Functional Measurement and Psychophysical Judgment. *Psychological Review, 77,* 153- 170.

Anderson, N. H. (1981). Foundation of information integration theory. New York: Academic Press.

Anderson, N. H. (1982). Methods of information integration theory. New York: Academic Press.

Beck, N., & Jackman, S. (1998). Beyond Linearity by Default: Generalized Additive Models. *American Journal of Political Science*, 42, 596-627

Berk, R. A. (2004). Regression analysis: A constructive critique: Advanced Quantitative Techniques in the social sciences series (volume 11). Sage Publications: Thousand Oaks, CA.

Bogen, J. & Woodward, J. (1988). Saving the phenomena. The philosophical Review, 97, 303-352.

Borda J.C. (1784). Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences.

Bousfield,W.A., & Sedgwick, C.H. (1944). An analysis of restricted associative responses. *The Journal of General Psychology*, *30*, 149–165.

Bröder, A. (2002). Take the best, Dawes' rule, and compensatory decision strategies: A regression based classification method. *Quality and Quantity*, *36*, 219–238.

Bröder, A. (2000). Assessing the empirical validity of the "take the best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1332–1346.

Bröder, A., & Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General*, *132*, 277–293.

Brunswik, E. (1939). The conceptual focus of some psychological systems. *Journal of Unified Science (Erkenntnis)*, 8, 36-49.

Brunswik, E. (1952). The conceptual framework of psychology. (International Encyclopedia of Unified Science, Volume 1, Number 10.) Chicago: The University of Chicago Press.

Brunswik, E. (1955), "Representative design and probabilistic theory in functional psychology," *Psychological Review*, 62, 193-217.

Brunswik, E., & Herma, H. (1951). Probability learning of perceptual cues in the establishment of a weight illusion. *Journal of Experimental Psychology*, 41, 281-290.

Burnham, K.P. and D.R. Anderson. 1998. Model Selection and Inference. Springer-Verlag, NY.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, 54(1), 149-185. doi:10.1111/j.1744-6570.2001.tb00090.x

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3), 494-509. doi:10.1037/0033-2909.114.3.494

Cliff, N. (1994). Predicting ordinal relations. *British Journal of Mathematical and Statistical Psychology*, 47(1), 127-150.

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31(3), 331-350. doi:10.1207/s15327906mbr3103_4

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

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

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVADM: A memory processes model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Dougherty, M. R., Franco-Watkins, A., & Thomas, R. P. (2008). The psychological plausibility of fast and frugal heuristics. *Psychological Review.115(1),* 199-211.

DuCharme, W. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology*, 85, 66-74

Dudycha, A. L. and Naylor, J. C.(1966). The Effect of Variations in the Cue R Matrix Upon the Obtained Policy Equations of Judges. *Education and Psychological Measurement,26,* 583-603.

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, 73(3), 221-230. doi:10.1037/h0028695

Einhorn, H. J., Komorita, S. S., & Rosen, B. (1972). Multidimensional models for the evaluation of political candidates. *Journal of Experimental Social Psychology*, 8(1), 58-73. doi:10.1016/0022-1031(72)90061-3

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128,* 309-331.

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, 112(4), 912-931. doi:10.1037/0033-295X.112.4.912

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101, 519-527.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, **83**, 37-64.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22, 1947-1975.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin and Review*, *5*, 370–389.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology *Psychological Review, 98*, 254–267.

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

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco - Watkins, and Thomas (2008). *Psychological Review, 115(1),* 230-237.

Gigerenzer, G., Hoffrage, U., & Kleinbo1ting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review,, 98,* 506-528.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L. J., Beatty, J., & Krueger, L. (1989). The empire of chance: How probability changed science and everyday life. Cambridge, UK: Cambridge University Press.

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

Goldberg, D. E. (2002). The design of innovation: Lessons from and for competent genetic algorithms. Addison-Wesley, Reading, MA.

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

Gonzalez, R. & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119,* 159-165.

Goodman, L. A. & Kruskal, W. H. (1979). Measures of association for cross classifications. New York: Springer-Verlag.

Griffiths, T. L. & Tenenbaum, J. B. (2006) Optimal predictions in everyday cognition. *Psychological Science 17(9),* 767-773.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211-244.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review,* 71(6), 438-456. doi:10.1037/h0040736

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438–456.

Hammond, K.R. (1980). Introduction to Brunswikian theory and methods. In K.R. Hammond and N.E. Wascoe (Eds.), *Realizations of Brunswik's experimental design*. San Francisco: Jossey-Bass.

Hartwig, M., & Bond, C. R. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4), 643-659. doi:10.1037/a0023589

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*, 494–508.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, **9**, 394-401.

Heider, F. (1958). The psychology of interpersonal relations. New York: John Wiley & Sons.

Herrnstein, R. J., & Heyman, G. M. (1979). Is matching compatible with reinforcement maximization on concurrent variable interval, variable ratio?. *Journal of the Experimental Analysis of Behavior*, 31(2), 209-223. doi:10.1901/jeab.1979.31-209

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review, 17*, 923-930.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2010). The central executive as a search process: Priming exploration and exploitation across domains. *Journal of Experimental Psychology: General*, 139(4), 590-609. doi:10.1037/a0020666

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116–131.

Ho, T.K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 1, 66-75.

Hogarth, R. M., & Karelaia, N. (2006). "Take-the-best" and other simple strategies: Why and when they work "well" with binary cues. *Theory and Decision*, 61, 205-249.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733-758.

Huber, P. J. (1981). Robust Statistics, Wiley.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-probability studies. *Psychological Review*, *71*, 42–60.

James, W. (1892). Psychology. Henry Holt and Company, New York.

Juslin, P., Winman, A., & Hansson, P. (2007). The *naïve* intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Revie*w, *114*, 678-703.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, XLVII, 263-291.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A metaanalysis of lens model studies. *Psychological Bulletin*, 134 (3), 404-426.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), Nebraska symposium on motivation. Lincoln: University of Nebraska Press.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28(2),* 107-128.

Kendall, M. (1938) A New Measure of Rank Correlation, *Biometrika*, 30, 81-89.

Kendall, M.G. (1970). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Kim, C., Yang, K., & Kim, J. (2008). Human decision-making behavior and modeling effects. *Decision Support Systems*, 45(3), 517-527. doi:10.1016/j.dss.2007.06.011

Lebreton, J., Burnham, K.P., Clobert, J., and Anderson, D.R. (1992) Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies. *Ecological Monographs*, 62, 67–118. doi:10.2307/2937171

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the “take the best” and the “rational” models, *Psychonomic Bulletin & Review*, 11, 2, 342 – 352.

Lichtenstein, S., P. Slovic, B. Fischhoff, M. Layman, and B. Combs. (1978) Judged frequency of lethal events. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 4, 551-578.

Long, J. D. (1999). A confidence interval for ordinal multiple regression weights. *Psychological Methods*, 4(3), 315-330. doi:10.1037/1082-989X.4.3.315

Long, J. D. (2005). Omnibus hypothesis testing in dominance-based ordinal multiple regression. *Psychological Methods*, 10(3), 329-351. doi:10.1037/1082-989X.10.3.329

Longo, M. R., & Lourenco, S. F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia, 45*, 1400-1407.

MacCallum, R. C., Cornelius, E. T., & Champney, T. (1979) Validity and cross-validity of metric and nonmetric multiple regression. *Applied Psychological Measurement*, 3, 463-468.

Magdon-Ismail, M., & Sill, J. (2008). A linear fit gets the correct monotonicity directions", Machine Learning, 70, 1, 21-43.

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

Micceri, T. (1989). The *unicorn*, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.

Myung, I. J. & Pitt, M. A. (2004). Model comparison methods . *Methods in Enzymology*, *383*, 351-366.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum. ISBN 0898590930

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best."*Organizational Behavior and Human Decision Processes*, *91*, 82–96.

Otto, A., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, 118(2), 274-279. doi:10.1016/j.cognition.2010.11.009

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1992). Behavioral decision research: A constructive processing perspective. Annual Review of Psychology, 43, 87-131.

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

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68 (1)*, 29-46.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-196.

Rupinski, M. T., & Dunlap, W. P. (1996). Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. *Educational and Psychological Measurement*, 56, 419-429.

Schmidt, F. L. (1995). Why all banding procedures in personnel selection are logically flawed. Human Performance. 8 (3), 165-177.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 2, 461–464.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*, 233–250.

Shanteau, J. (1974). Component processes in risky decision making. Journal of Experimental Psychology, 103, 680-691.

Shanteau, J. & Thomas, R. P. (2001) Fast and frugal heuristics: What about unfriendly environments. *Behavioral & Brain Science*, *22*, 5.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, *27*, 125-140.

Shepard. R. (1966). Metric structures in ordinal data, *Journal of Mathematical Psychology*, *3*, 287-3J5.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 2, 129-138.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (Ed.), *Architectures for intelligence.* (pp. 15–39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Steinmann, D. O., & Doherty, M. E. (1972). A lens model analysis of a bookbag and poker chip experiment: A methodological note. *Organizational Behavior and Human Performance*, *8*, 450–455.

Stewart, N., Chater, N. & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology.*

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10(7),* 309-318.

Thomas, R. P., & Dougherty, M. R. P., Sprenger, A., & Harbison, I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*. 115(1), 155-185.

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in magnitude estimation. *Journal of Experimental Child Psychology, 101,* 20-51.

Trujillo-Ortiz, A. & Hernandez-Walls, R. (2003). Mskekur: Mardia's multivariate skewness and kurtosis coefficients and its hypotheses testing. A MATLAB file. [WWW document]. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=3519

Trujillo-Ortiz, A., Hernandez-Walls, R., Barba-Rojo, K.,& Cupul-Magana, L. (2007). HZmvntest:Henze-Zirkler's Multivariate Normality Test. A MATLAB file. [WWW document]. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17931

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch and by Hammond, Hursch, and Todd. *Psychological Review*, *71*, 528–530.

White, C. M., & Koehler, D. J. (2007). Choice strategies in multiple-cue probability learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 757-768. doi:10.1037/0278-7393.33.4.757

York, K. M., Doherty, M. E., & Kamouri, J. (1987). The influence of cue unreliability in a multiple cue probability learning task. *Organizational Behavior and Human Decision Processes*, *39*, 303–317.

Young, F. M., De Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. Psychometrika, 41, 505-528.1977-11837-001.

Authors note

Michael R. Dougherty, Department of Psychology, University of Maryland, College Park, MD 20742; Rick Thomas, Department of Psychology, University of Oklahoma, Norman, OK, 73019.
Both authors contributed equally to this paper, and authorship order is arbitrary. The research presented in this paper was partially supported by a grant SES-0620062 to MRD. The authors thank Randy Engle and Lael Schooler for providing their data sets, and Erica Yu, Amber Sprenger, Louis Narens, Jorge Mendoza, Paul Hanges, Gerd Gigerenzer, J. Brenden Richie, and Will Beasley for their many valuable comments. Correspondance regarding this article can be addressed to mdougher@umd.edu or Rickey.P.Thomas-1@ou.edu.

Figure 1. Hit and false alarm rates as a function of sample size for GeMM, OLS-BIC, and OLS-Wald for a Linear environment.

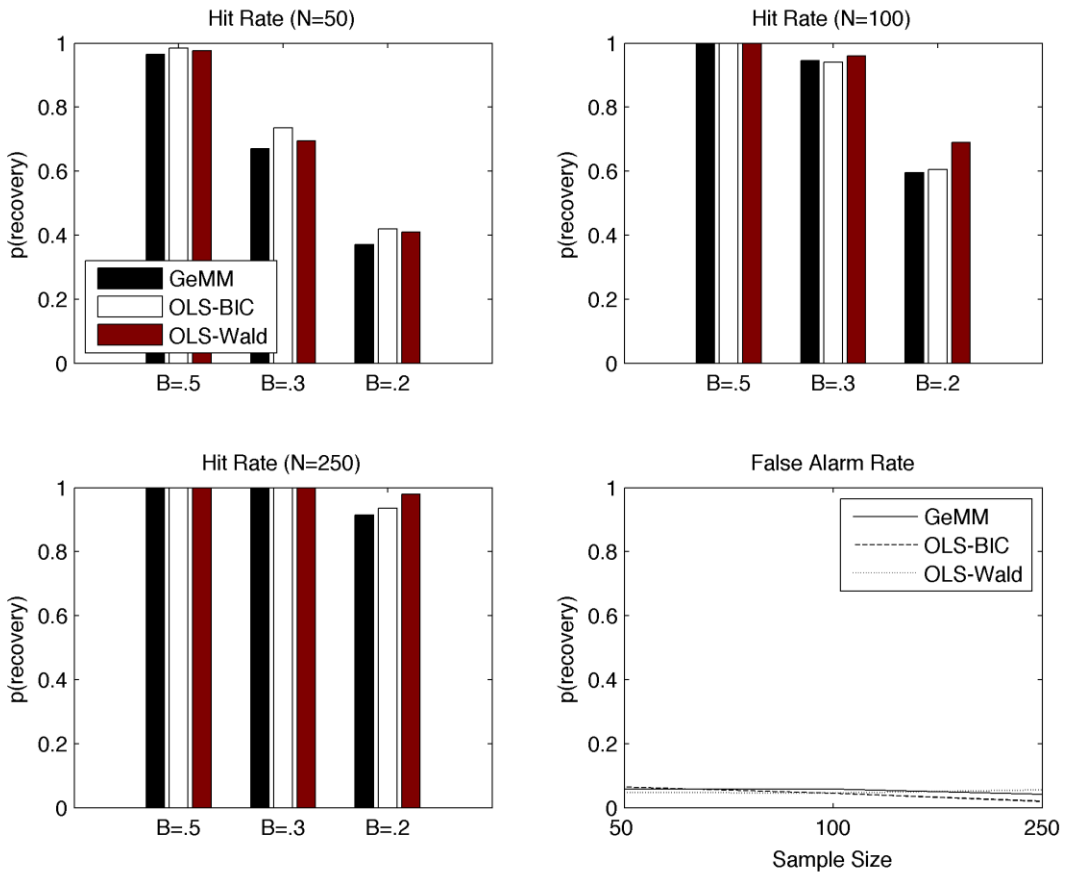


Figure 2. Hit and false alarm rates as a function of sample size for GeMM, OLS-BIC, and OLS-Wald for a Non-linear environment.

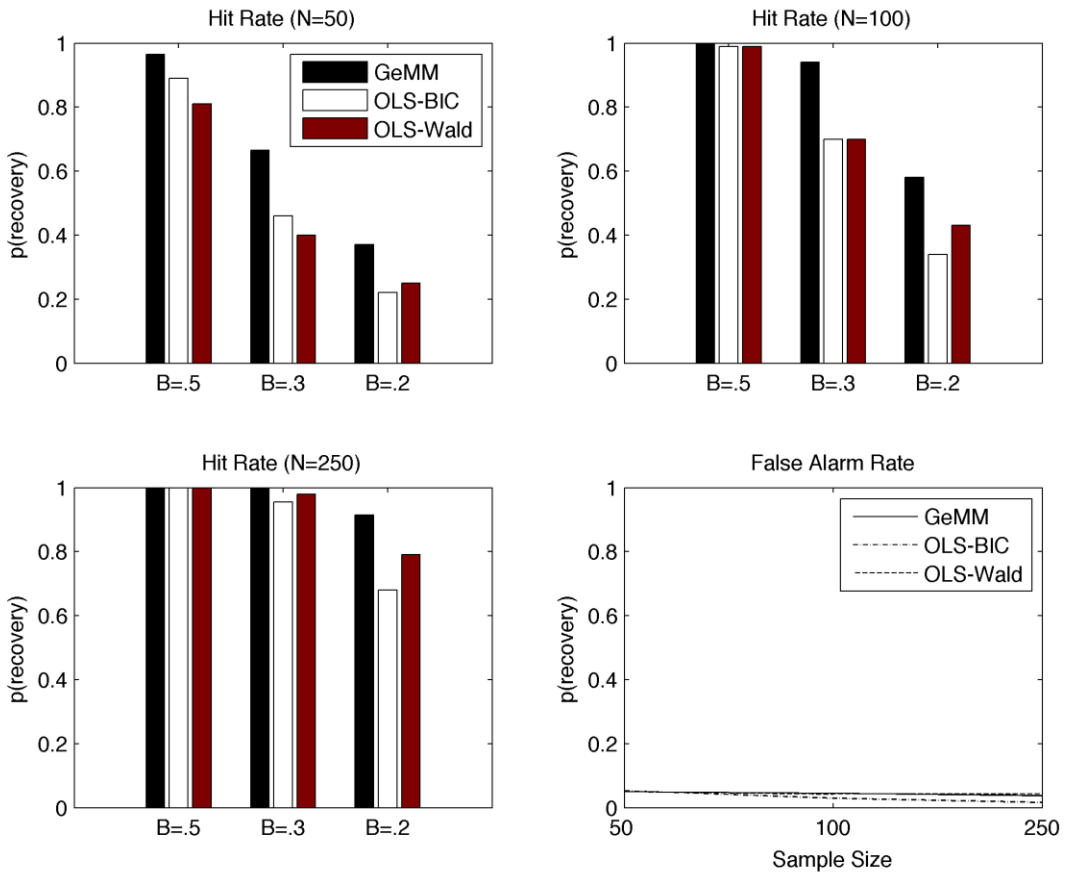
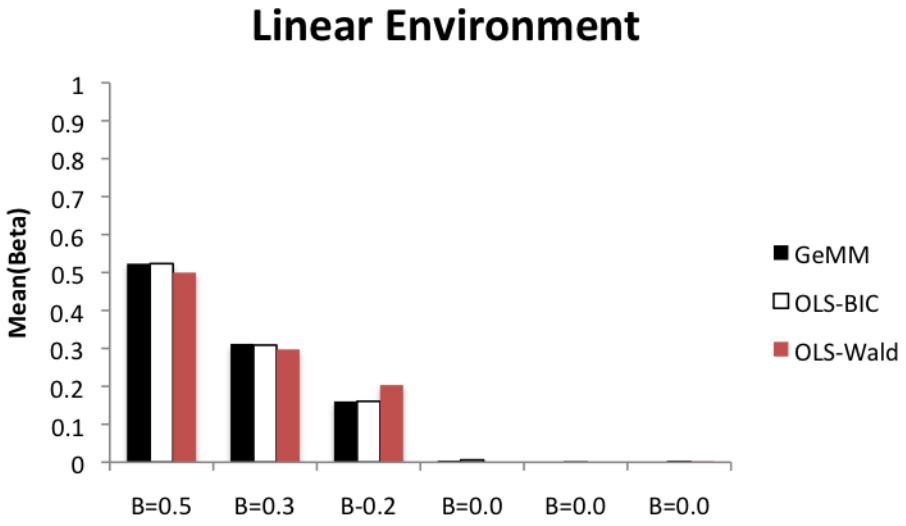


Figure 3. Mean normalized parameter estimates of recovered parameters for GeMM, OLS-NHST, and OLS-BIC in the linear and non-linear environments for sample size N=100.



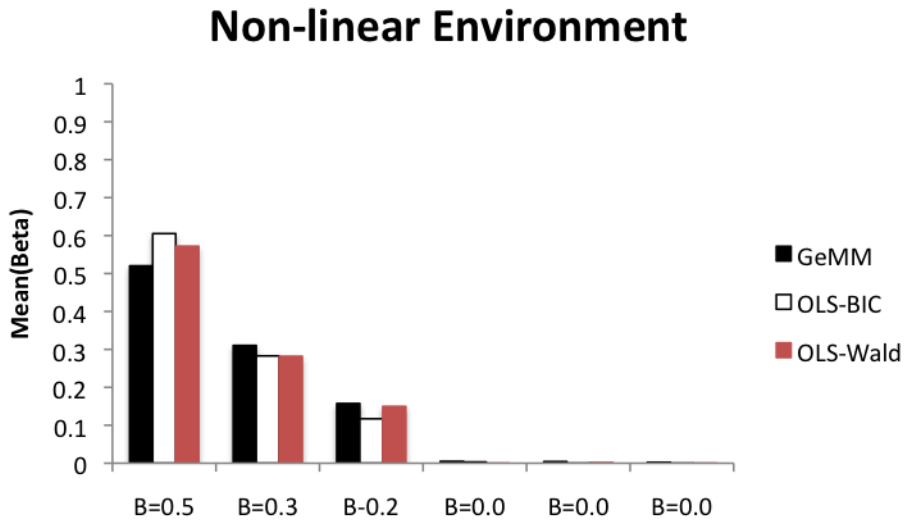


Figure 4. Proportion of concordances for the estimation and crossvalidation sample in the linear and non-linear environments for sample size N=100.

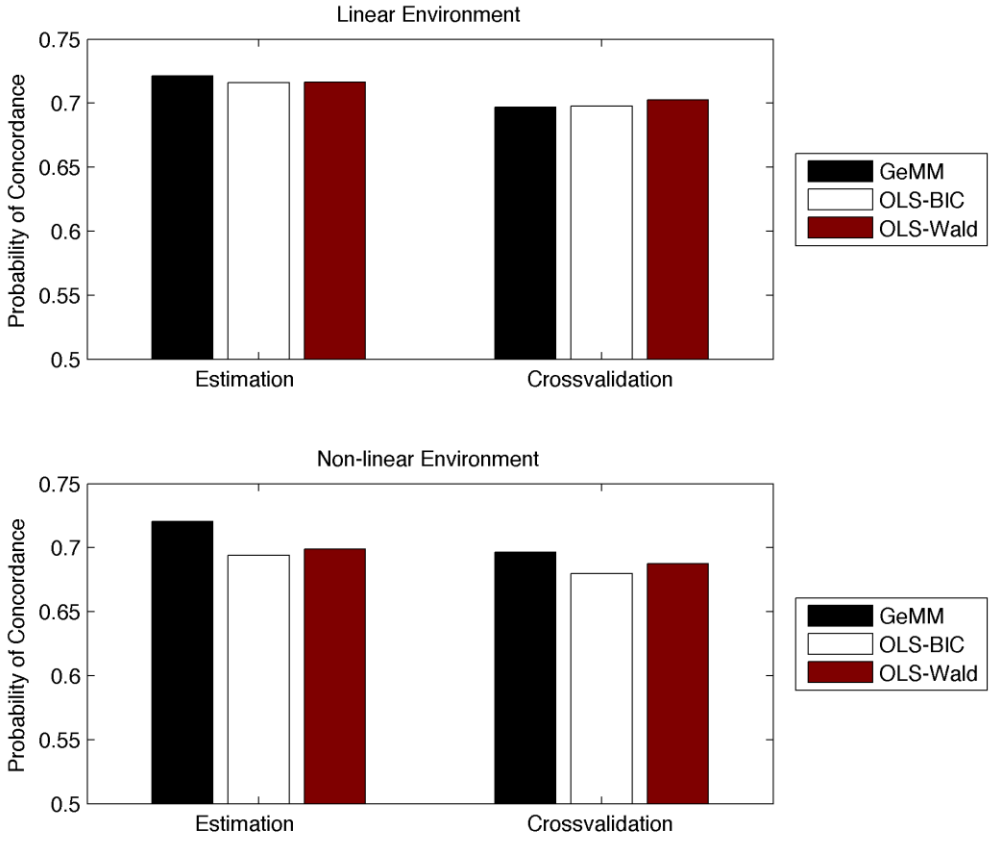


Figure 5. Mean Kendall tau correlations for the estimation and crossvalidation sample in the linear and non-linear environments for sample size N=100.

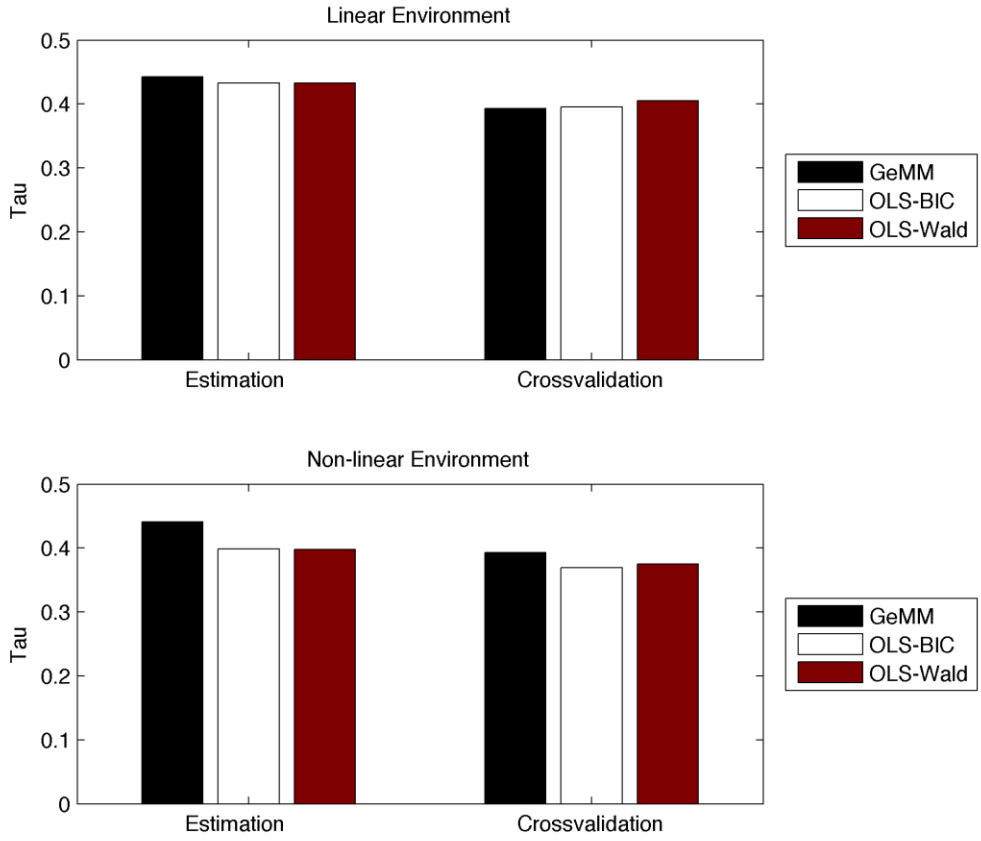


Figure 6.

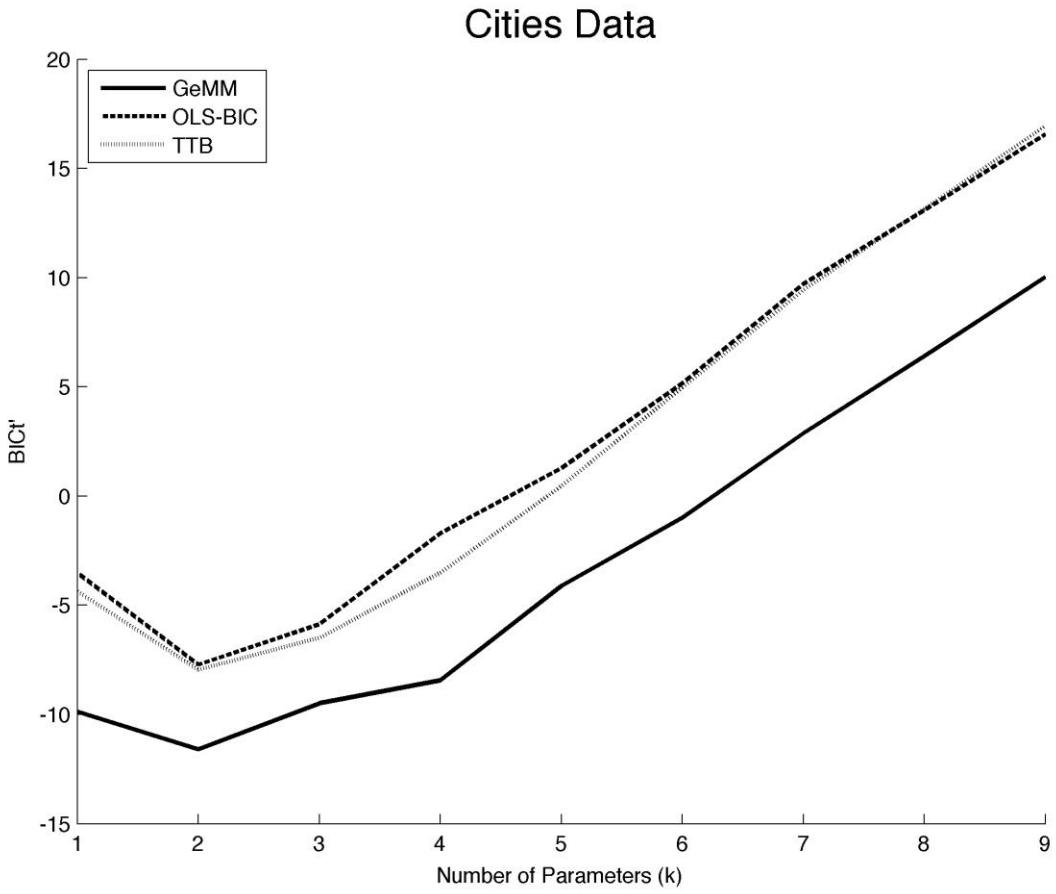

Cities Data
GeMM
OLS-BIC
TTB
BICt'
Number of Parameters (k)
20
15
10
5
0
-5
-10
-15
1
2
3
4
5
6
7
8
9

Figure 7. Probability of concordance for the estimation and cross-validation sample for the Cities Data.

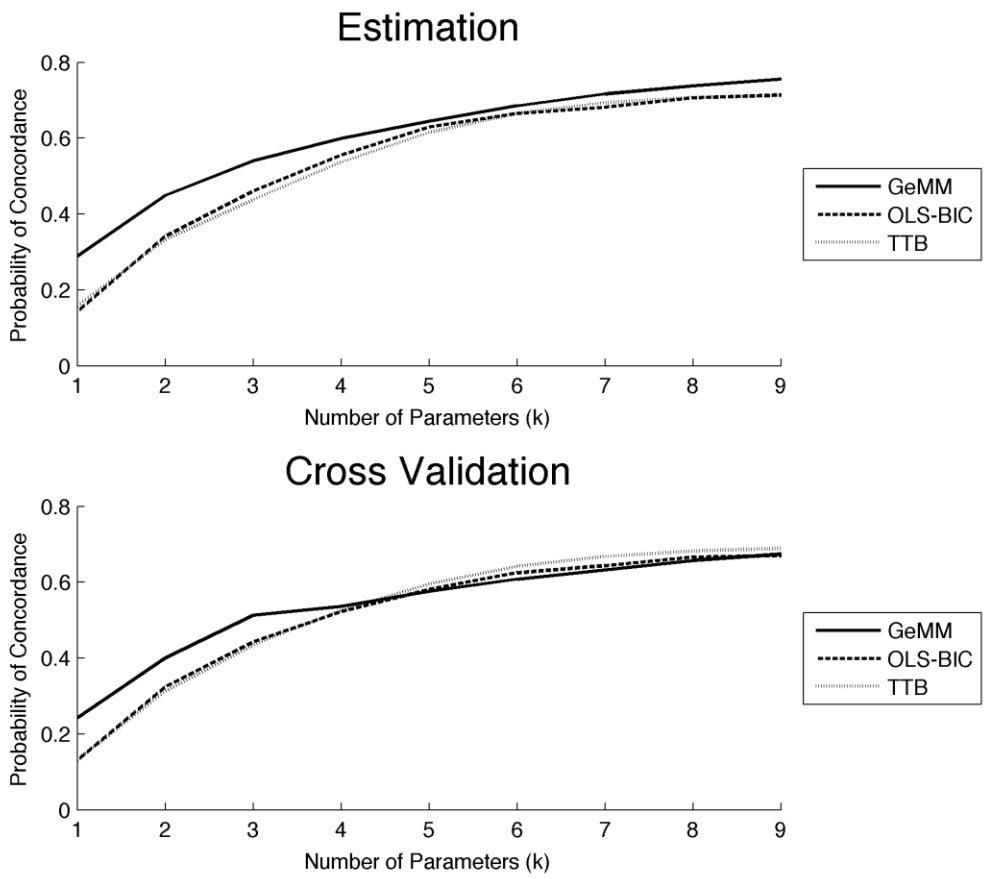


Figure 8. Mean Kendall tau for the estimation and cross-validation sample for the Cities Data.

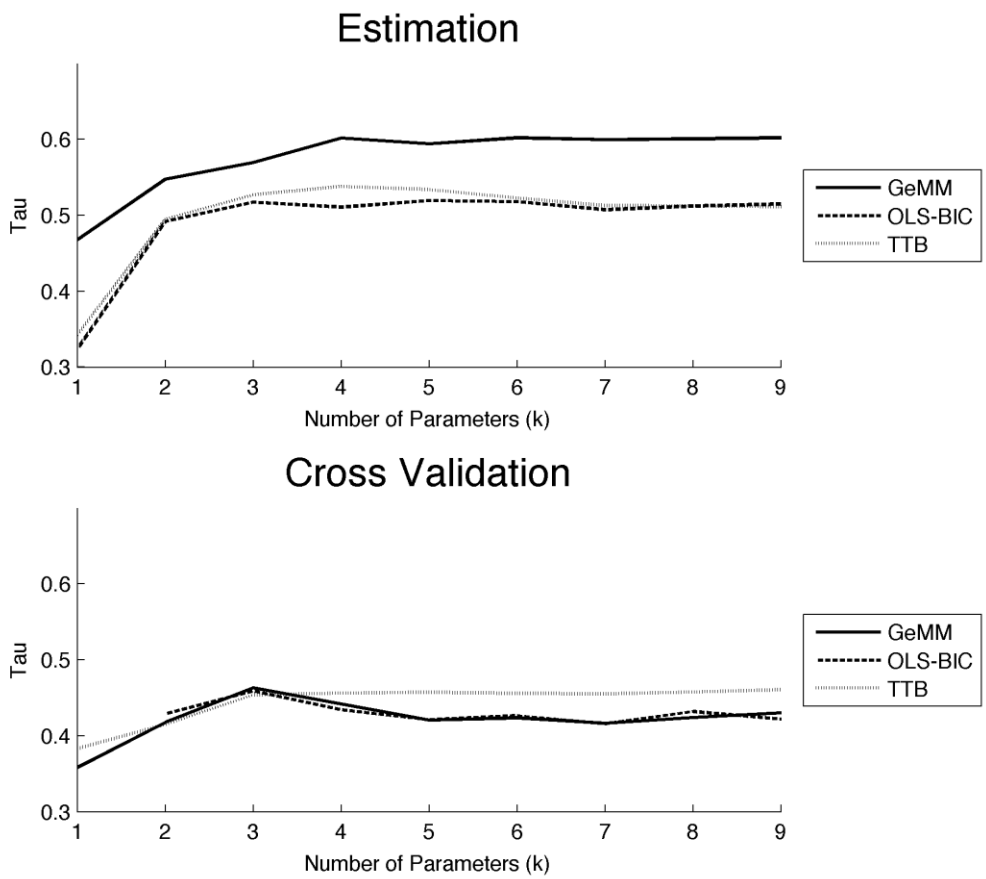


Figure 9. **Mean BICt` plotted as a function of the number of parameters (k) for each strategy for the Salary Data.**

Professors Salary Data

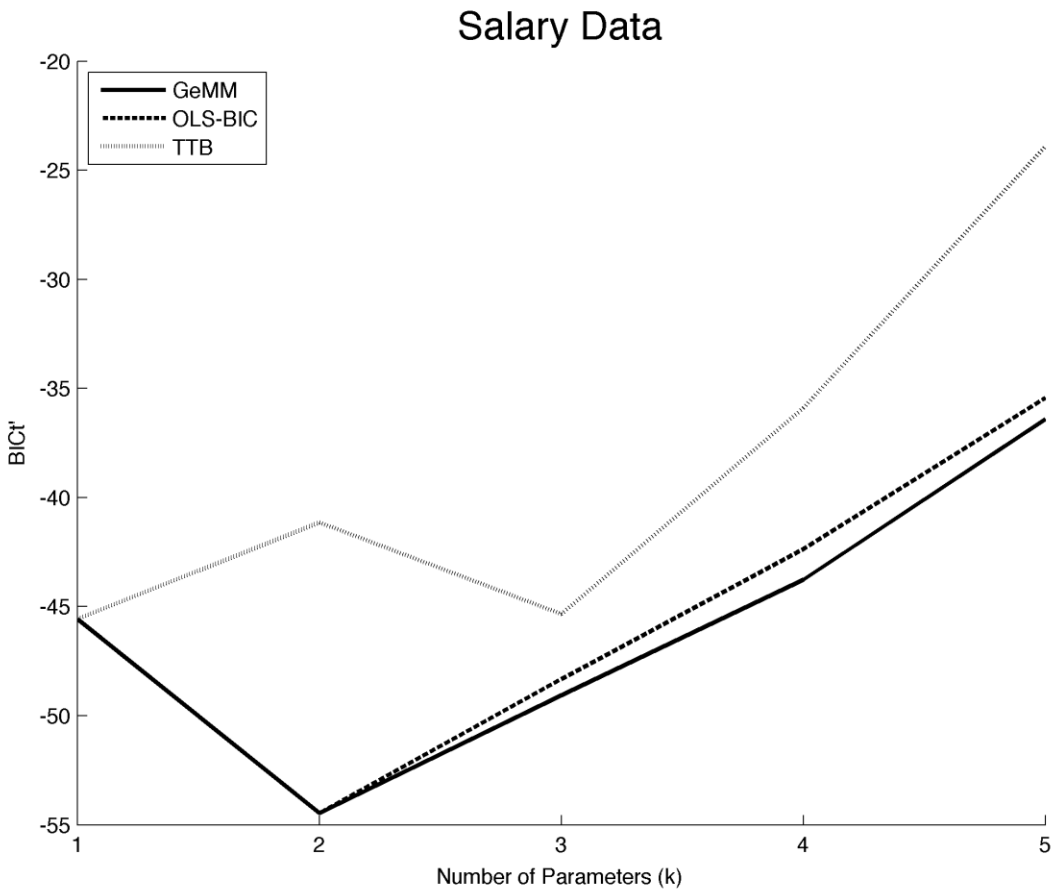


Figure 10. **Mean concordances (top panel) and mean Tau's (bottom panel) plotted as a function of the number of parameters (k) for each strategy for the Salary Data.**

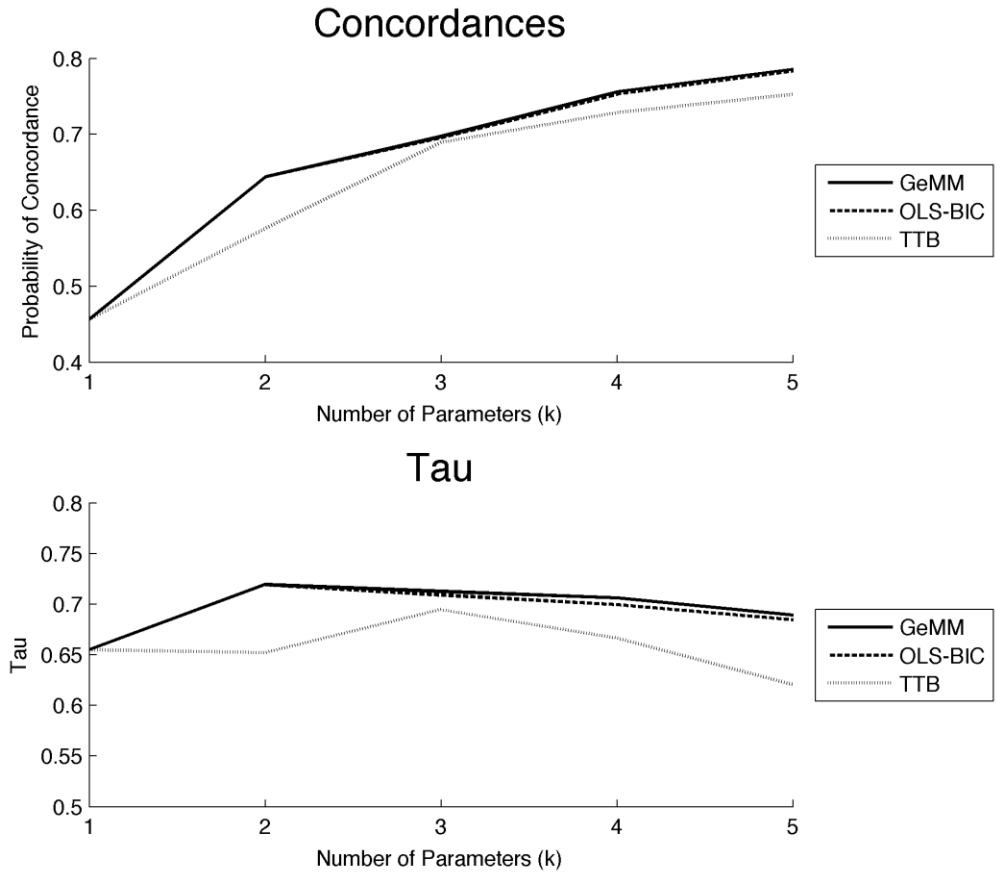


Figure 11. Illustration of the classic Lens model framework. The Lens model parameters illustrated are described in Table 3.

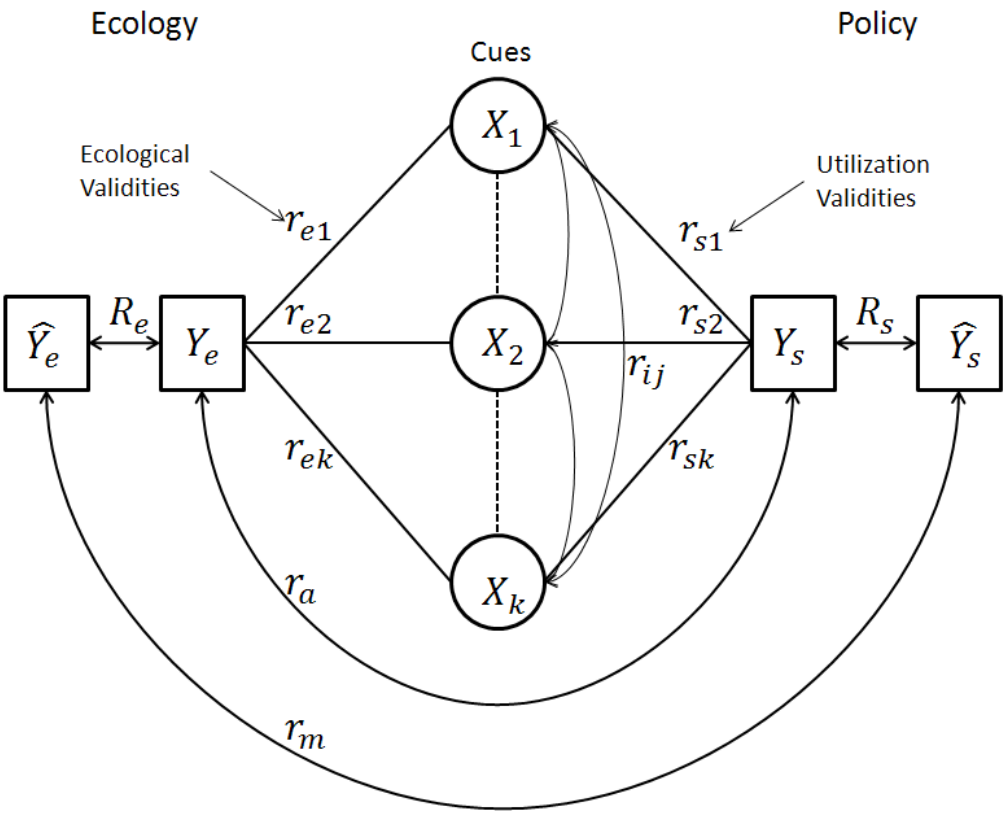


Table 1. Probability of recovering each parameter off of N/2 observations.

	O-span	R-span	V-SAT	Q-SAT	B-Span	F-Span	K-Track	C-Span	Cattell
GeMM	0.04	0.075	0.075	0.865	0.02	0	0.095	0.085	0.965
OLS-BIC	0.05	0.035	0.03	0.71	0.06	0.005	0.135	0.095	1
OLS-Wald	0	0.05	0	0.435	0.02	0.07	0.045	0.05	1
RLS Wald (Huber)	0	0.05	0.005	0.575	0.005	0.045	0.02	0.045	1
RLS Wald (BiSquare)	0	0.07	0.01	0.585	0	0.04	0.015	0.055	0.985
Bayes-Wald	0	0.028	0	0.389	0.011	0.072	0.038	0.022	1
Ridge	1	1	1	1	1	1	1	1	1

Table 2. Fit indices and cross-validation statistics for all seven regression procedures.

		BICt'	BIC	P(c)	Tau	r	$\bar{k}$
GeMM	Estimation	-32.4973	-40.4052	0.7422	0.5633	0.7255	2.22
	Cross-Validation			0.7148	0.5071	0.6989	
OLS-BIC	Estimation	-29.6933	-42.7902	0.7312	0.5432	0.7305	2.12
	Cross-Validation			0.7122	0.5042	0.6998	
OLS-Wald	Estimation	-25.7679	-40.9317	0.6998	0.5108	0.7158	1.67
	Cross Validation			0.6875	0.4861	0.6778	
Huber-RLS Wald	Estimation	-26.9341	-41.4593	0.7093	0.5214	0.7200	1.74
	Cross-Validation			0.6981	0.4988	0.6876	
Bi-Square-RLS Wald	Estimation	-27.028	-41.1566	0.7094	0.5227	0.7188	1.76
	Cross-Validation			0.6985	0.5002	0.6876	
Bayes-OLS Wald	Estimation	-27.0316	-41.3713	.6943	.5060	.7144	1.56
	Cross-Validation			.6856	.4875	.6821	
Ridge Regression	Estimation	-1.427	-9.7332	.7368	.5472	.7144	9
	Cross-Validation			.7195	.5120	.6973	

p(c) = percent concordance, Tau = Kendall's tau correlation, r = Pearson's r correlation, k = number of parameters in model, BIC=Bayesian Information Criterion, BICt' = Bayesian Information Criterion based on the weighted tau to r transformation,

Table 3. Descriptions of the Lens Model Components in the original linear model and recast within GeMM.

Linear Lens Model Component	**Standard Description**	**Monotone Lens Model Component**	**Monotone Description**
X_*	Cues are assumed to be metric.	X_*	Cues are assumed to be monotone.
r_{S*}	Best fit linear regression coefficients for each cue regressed on the judge's evaluations. The number of statistically significant cues is often interpreted as the frugality or complexity of the judge's policy and the magnitude of the linear regression coefficients are interpreted as the extent to which the cues were 'utilized' in judge's policy.	r_{S*}^{τ}	Best fit monotone regression coefficients, estimated via GeMM for each cue regressed on the judge's evaluations. The number of cues in the best fit model, using BIC_{τ} to trade fit and complexity, can be interpreted as the frugality of the judge's policy and the magnitude of the monotone regression coefficients are interpreted as the extent to which the cues were 'utilized' in the judge's policy.
r_{e*}	Best fit linear coefficients for each cue regressed on the criterion. The statistical significance and the magnitude of the linear coefficients are interpreted as the 'validity' of the cues to predict the criterion variable.	r_{e*}^{τ}	Best fit monotone regression coefficients, estimated via GeMM for each cue regressed on the criterion. The number of cues in the best fit model, using BIC_{τ} to trade fit and complexity, and the magnitude of the monotone regression coefficients are interpreted as the 'validity' of the cues to predict the criterion variable.
Y_S	Judgments are assumed to be metric.	Y_S	Judgments are assumed to be monotone.
Y_e	The criterion variable is assumed to be metric.	Y_e	The criterion is assumed to be monotone.

$\widehat{Y}_s$	The predicted values of the best-fit linear model of the cues regressed on the judge's judgments.	$\widehat{Y}_s^{\tau}$	The predicted values of the best-fit GeMM model of the cues regressed on the judge's judgments
$\widehat{Y}_e$	The predicted values of the best-fit linear model of the cues regressed on the criterion	$\widehat{Y}_e^{\tau}$	The predicted values of the best-fit GeMM model of the cues regressed on the criterion
R_s	Coefficient of determination of the best-fit linear model of the cues regressed on the judge's judgments. Also, the Pearson correlation coefficient between the predicted values of the judge's judgments, estimated via linear regression, and the observed judgments. The metric is often interpreted as the ability or fit of the linear model to "capture" the judgment process.	R_s^{τ}	The Kendall's tau estimate of the Pearson product moment correlation coefficient, r, between the predicted values of the judge's judgments, estimated via GeMM regression, and the observed judgments. The metric can be interpreted interpreted as the ability or fit of the monotone model to "capture" the judgment process.
R_e	Coefficient of determination of the best-fit linear model of the cues regressed on the criterion. Also, the Pearson correlation coefficient between the predicted values of the criterion, estimated via linear regression, and the observed criterion. The metric is often interpreted as the ability or fit of the linear model to "capture" the ecology.	R_e^{τ}	The Kendall's tau estimate of the Pearson product moment correlation coefficient, r, between the predicted values of the criterion, estimated via GeMM regression, and the observed criteria. The metric can be interpreted as the ability or fit of the monotone model to "capture" the ecology.
r_a	The Pearson correlation between the judge's judgments and the observed criterion. The measure is interpreted as the quality or accuracy of the judge to predict the criterion values in terms of variance-accounted-for.	r_a^{τ}	The Kendall's tau estimate of the Pearson product moment correlation coefficient r, via GeMM regression, between the criteria and the judge's judgments. The measure can be interpreted as the quality or accuracy of the judge to predict the criterion values.
r_m	Often referred to as match, the correlation between the predicted values of the judge's judgments and the predicted values of the criterion.	r_m^{τ}	The monotone surrogate of match, the Kendall's tau estimate of the Pearson correlation between the predicted values of the judge's

			judgments and the predicted values of the criterion estimated via GeMM.
r_{ij}	The inter-cue correlations between the predictors or cues.	r_{ij}^{τ}	The Kendall's tau estimate of the Pearson product moment correlation coefficient, r, between the predictors or cues. In process terms, cues that are highly correlated are assumed to substitute for each other --a process referred to as vicarious functioning.

Table 4. Decision models and relation to GeMM.

Model	Description	Number of parameters (K)	Model selection procedure
Take the best	Define all possible 1-parameter models. Order models according to predictive validity. For each paired comparison, search models from best to worst until one discriminates between the pair of options.	K = Number of models required to discriminate all possible dyads.	None. Use all possible models sequentially.
Take the Best-select	Define all possible 1-parameter models. Order models according to predictive validity. Select subset of models that minimize BIC. For each paired comparison, search selected models from best to worst until one discriminates between the pair of options.	K = Number of models required to minimize the BICt'.	Estimate tau based on concordances and disconcordances. Transform tau to r, use equation 18 or any other model selection procedure that trades parsimony for fit
Minimalist	Define all possible 1-parameter models. For each paired comparison, randomly choose model without until one discriminates between the pair of options.	K = Number of models required to discriminate all possible dyads.	None
Minimalist - Select	Define all possible 1-parameter models. Select subset of	K = Number of models required to minimize the BICt'	Estimate tau based on concordances and

	models that minimize BIC. For each paired comparison, randomly choose model from selected set until one discriminates between the pair of options.		disconcordances. Transform tau to r, use equation 18 or any other model selection procedure that trades parsimony for fit.
Franklin's Rule (Weighted additive model)	Define weight vector. Sum product of weight vector by data vector. *Multiple regression is the special case where weight vector is optimized my minimizing least squares.	K= total number of variables in dataset	
Franklin's Rule - Select	Define weight vector. Sum product of weight vector by data vector. *Multiple regression is the special case where weight vector is optimized my minimizing least squares.	K = Number of variables included in the model with the lowest BICt'.	Estimate tau based on concordances and disconcordances. Transform tau to r, use equation 18 or any other model selection procedure that trades parsimony for fit.
Dawes' Rule (Unit weight additive model)	Set weight vector to unit values (1,0,-1). Sum product of weight vector by data vector	K= total number of variables in dataset	None. Use all predictor variables simultaneously.
Dawes' Rule - Select	Set weight vectors to unit values. Sum product of weight vector by data vector	K = Number of variables included in the model with the lowest BICt'.	Estimate tau based on concordances and disconcordances. Transform tau to r, use equation 18
Borda count	Set weight vectors to unit values. Rank transform the predictors. Sum product of weight	K = Number of predictors (voters) in the dataset	None. Use all predictor variables simultaneously.

	vector by data vector		
TTB-GeMM	Define all possible K>=1 parameter models. Order models according to BICt'. For each paired comparison, search models from best to worst until one discriminates between the pair of options.	-	-
Minimalist-GeMM	Define all possible K>=1 parameter models. For each paired comparison, randomly choose model without until one discriminates between the pair of options.	-	-

Appendix A

The model competition for modeling the Engle et al. (1999) cognitive ability data included GeMM, OLS-BIC, OLS-Wald, and four additional techniques including two versions of robust regression, one using the Huber weighting function and a version using the bisquare weighting function, Bayesian regression with normal priors, and ridge regression. Robust regression is a technique used in contexts for reducing the impact of extreme scores, and is therefore less sensitive to the presence of outliers. Both robust procedures were implemented within MatLab using a WALD test. The Bayesian and ridge regression procedures were implemented in SAS, with code written by the authors. The Bayesian model used normal priors with variable selection based on the Wald 95% credible interval. We implemented Ridge Regression, which is a variant of OLS that applies a shrinkage penalty, Lambda, to the predictors. Although this penalty slightly biases the parameter estimates, it can lead to substantial decreases in their variance when the predictors are highly collinear. We employed a variant of the BIC (see equation A1) to select the Lambda for each estimation sample that optimally traded fit (RSS) with a measure of flexibility (effective df). The parameter estimates associated with the best-fit Lambda were then applied to the cross-validation sample.

$$BIC_{\lambda} = \log RSS_{\lambda} + df_{\lambda} / n \log n, \text{ where } df_{\lambda} = \text{trace}[X(X^tX + \lambda I)^{-1}X^t] \qquad \text{eq. A1}$$

For all procedures, we randomly sampled 50% of the total sample to form an estimation sample, and used the remaining 50% as the hold out, or cross-validation sample. For each run, we estimated the model on the estimation sample then applied the statistical model to predict observations in the cross validation sample. This was repeated 200 times for each approach. For the sake of comparison, we used the model based on the full sample as the criterion for

evaluating statistical power. Using the full sample, five approaches (OLS-BIC, OLS-Wald, Huber-RLS, Bi-Square RLS, and Bayesian-Wald identified a two-parameter model consisting of Q-SAT and Cattell's culture fair test as the best-fit model for predicting Raven's Progressive Matrices. Ridge regression is not typically implemented within a variable selection algorithm, so it was implemented by including all 9 predictors. Table 1 presents the fit indices from the cross-validation exercise. Table 2 presents the probability that each parameter was recovered for N/2.