Judgment and Decision Making, Vol. 3, No. 8, December 2008, pp. 667–678

On the complexity of traffic judges' decisions

David Leiser* & Dov-Ron Schatzberg
Ben-Gurion University of the Negev

Abstract

Professional judges in traffic courts sentence many hundreds of offenders per year. Using 639 case files from archives, we compared the Matching Heuristic (MH) to compensatory, weighing algorithms (WM). We modeled and cross validated the models on different subsets of the data, and took several other methodological precautions such as allowing each model to select the optimal number of variables and ordering and weighing the variables in accordance to different logics. We did not reproduce the finding by Dhami (2003), who found the MH to be superior to a compensatory algorithm in modeling bail-granting decisions. These simulations brought out the inner logic of the two family of models, showing what combination of parameters works best. It remains remarkable that using only a fraction of the variables and combining them non-compensatorily, MH obtained nearly as good a fit as the weighing method.

Keywords: bounded rationality; judgments; frugal; take-the-best; decision-making; simple heuristics; models; matching heuristic.

1 Introduction

This paper compares compensatory and non-conpensatory models as descriptions of how decisions are made in real-life settings by experienced practitioners. The study concerns traffic judges, and the punishment they mete out to alleged offenders brought in their court. Judges in traffic courts are professional decision makers. Their decisions are, for them, routine, and they handle dozens of cases every month. For the defendants appearing in court, the decisions are significant, involving sometimes hefty fines, driving license suspension or even prison. We asked how many elements traffic judges consider and how complex their decision patterns are. Such a study, relying on actual cases of professional decision-making, contributes to the debate about the plausibilitiy of the fast and frugal heuristics as psychological models of cognition (Bröder & Newell, 2008; Dougherty et al. 2008; Gigerenzer et al., 2008).

There is an extensive literature on how decision makers proceed when they have to rely on multiple cues to come to a decision. Many researchers maintain that decisions makers do not rely on all the information available, or combine it in any sophisticated way. Instead, it is claimed, they rely on simple heuristics that sometimes turn out to be surprisingly efficient and as valid as the more complex approaches favored by a rational analysis (Dhami, 2003; Gigerenzer et al. 1999, Smith & Gilhooly, 2006). This paper analyzes and attempts to model decisions made in real life by experienced traffic judges. How simple or how complex are the judgments meted out? Do traffic judges consider all the elements of the case? To what extent do they integrate all the information available to them?

There are at least two causes for the superficial treatment of available information. The first is cognitive limitation. People do not use an optimal approach because the proper integration of all the information available is beyond their mental powers. All humans have limited cognitive abilities (e.g., Cowan, 2001; Halford, Wilson, & Phillips, 1998), and this fundamental difficulty is exacerbated by the conditions under which traffic court judges have to work. Their workload is high, with only a few minutes available for a case, each characterized by some twenty parameters. Strategy selection is contingent on task demands (Payne, Bettman, & Johnson, 1993). Time pressure, and specifically the time available per variable has been shown to affect decision making (Balzer, Doherty, & O'Connor, 1989; Chewning & Harrell, 1990; Eppler & Mengis, 2004; Lee & Lee, 2004;) resulting in fewer variables being considered and the replacement of complex cognitive strategies by simpler ones (Weenig & Maarleveld, 2002; but see also Bröder & Newell, 2008, who qualify these findings)

Processing hundreds of cases month after month, traffic court judges acquire a vast experience. There is there-

*We gratefully acknowledge the contributions of Terri Yaskil and Orni Pachman and the assistance of Lilach Temelman and Tamar Lin. Helpful comments were contributed by members of the Center for Decision Making and Economic Psychology, Ben-Gurion University. We also thank Arndt Bröder, Jon Baron and anonymous reviewers for their observations and suggestions. Address: David Leiser, Dept of Psychology, Ben Gurion University PO Box 632, 84105 Israel. Email: dleiser@bgu.ac.il.

fore reason to expect their decisions to settle into some pattern. However, and this is the second factor, they do not have the benefit of feedback, and this has important consequences. Feedback on the appropriateness of one's decisions has been shown to be essential to developing good strategies (Balzer et al., 1989; Chenoweth, Dowling, & Louis, 2004; Youmans & Stone, 2005). In particular, while interactions are often ignored in multivariable decisions (Dawes, 1979) this is not the case when decision makers have habitual access to feedback on their decisions and care greatly about the outcome (Bröder, 2003; Ceci & Liker, 1986). Even if the judges knew what happened to the offenders they sentenced, they lack a clear criterion by which to evaluate the quality of their decision. Judges have little incentive to develop complex decision rules. Yet they surely do not render random judgments. Instead of accuracy, they may strive for an adequate judgment, one that they consider appropriate to the circumstances and expresses their attitudes about each case. Yet even if this criterion of *adequacy* replaces *accuracy*, it is unlikely that the judges attempt to reach the best possible decision in every case. Since their task is routine and each case must be quickly dispatched lest the backlog increase, it is plausible they do not invest as much care as judges in higher courts, where stakes arehigher and acquaintance with the offender protracted. They may also use methods that are less well thought out than if resources were unlimited, motivation high and information complete.

1.1 Fast and Frugal or Compensatory

The Fast and Frugal approach to decision-making holds that the use of simpler heuristics does not necessarily exact a cost. The heuristics are purported to provide psychologically plausible cognitive process models that describe a variety of judgment behaviors and enable the user to draw inferences or make decisions with a minimum of knowledge and computational effort. Among those heuristics, Take The Best (TTB) (about which more below) is the most widely studied (Dougherty, Franco-Watkins, & Thomas, 2008). Based on extensive modeling, its proponents claim that, under appropriate conditions, such heuristics can be as accurate as weighted linear models, even though they do not exploit all the information available and are non-compensatory (Czerlinski et al., 1999; Gigerenzer & Goldstein, 1999). Arguments and analyses related to TTB apply equally to related but less studied models, such as the Matching Heuristic (MH), the model that we will examine in detail here.

Although accuracy is not a relevant goal for judges, the latter may well use such heuristics as a way to achieve adequacy. The exact relation between TTB and ecological validity recently came under scrutiny (Dougherty, Franco-Watkins, et al., 2008; Dougherty, Thomas, & Franco-Watkins, 2008; Gigerenzer, Hoffrage, & Goldstein, 2008). The conclusion of this clarification process seems to be that TTB assumes a *subjective* rank order of cues (see Gigerenzer & Goldstein, 1996, p. 653), not an order in terms of their ecological validities (Gigerenzer, Hoffrage, & Goldstein, 2008). The output of TTB may then sometimes be seen as expressing subjective preferences, rather than judgments about the best objective decision. We mentioned the time pressure under which judges labor. What this means for strategy choice requires careful analysis, as there are at least two ways in which efficiency considerations may affect it: the choice of a non-compensatory rather than a compensatory strategy, and the use of selected rather than all the variables. TTB proceeds in a lexicographic, successive approach. One variable is considered at a time, and its indication is followed when it is clear enough. Only if the first variable fails to point to a conclusion, is the next variable considered, then the next, and so forth. Thus, TTB does not use all the variables available, nor does it combine in a compensatory way those that it uses. Bröder & Newell (2008) observe in their extensive review that the cognitive costs of running a compensatory strategy may have been overestimated, while the principal cost lies in "information search" (time pressure, memory retrieval, etc.).

Relevant laboratory studies on this issue use a paradigm derived from category learning. Participants learn to make correct judgments for a set of real-world stimuli based on feedback, and are then asked to make additional judgments (without feedback) for cases in which the TTB and the competing model made different predictions. Lee and Cummins (2004) compared TTB and the "rational approach" that terminates only when all available information has been assessed, and found inter-individual differences in the strategy used. However, the relevance of such studies is limited by the role of extensive practice, which recent studies show make a major difference, and by the absence of feedback in our case. Nosofsky and Bergert (2007; see also Bergert & Nosofsky, 2007) found that behavior changed markedly when overlearning takes place. Performance "skyrocketed," and detailed modeling analyses of their RT data suggest that this success was due to subjects learning to recode the correlated attributes into "higher order configural cues." They then could engage in a series of rule-based tests involving the recoded cues, "much in the spirit of a TTB process" (see Garcia-Retamero, Hoffrage, Dieckmann, & Ramos, 2007).

In a study on which we modeled ours to some extent, Dhami and Ayton (2001) had magistrates evaluate a set of realistic cases and for each decide whether a defendant should be granted bail. They then compared the

decisions made by magistrates to two weighted linear models that integrated all variables, and to the Matching Heuristic (based on principles similar to the TTB heuristic — see below for its detailed specification). Dhami's work (Dhami,2003; Dhami & Ayton, 2001) was criticized on methodological grounds by Bröder (2002; Bröder & Schiffer, 2003a, 2003b), who pointed out that the way the competition between the models was set up by Dhami and Ayton was unfair, because the matching heuristic had one extra free parameter. We have considered their criticism as well as several additional ones in designing our study. We view the matching heuristic (MH) and the linear weighed model (WM) as two families of models, with several possible realizations defined by the settings of several design parameters. We varied these parameters systematically. It is only after selecting the best representative of each of the two families of models, the ones with the best combination of parameters, that we will be able to decide on the relative merits of the two approaches, compensatory or non-compensatory, to modelling actual decisions by traffic court judges.

1.2 Modeling approach

We will attempt to determine whether traffic judges' are well fit by a model like TTB, whether they use a compensatory strategy, how much information they use, and how complex their decision pattern is. This will be done by a modeling approach. We will also use cross-validation of subsets of archival data to ensure reliable findings. Roberts & Pashler (2000; see also Rodgers & Rowe, 2002), discussing the widespread use of "model-fitting" in psychological science, have argued that a good fit of a model, defined as a high degree of successful data reconstruction, does not in itself guarantee a sound inference about the validity of the model.[1] Such an inference is valid only if the model's fit is clearly superior to the fit of other reasonable competing models of comparable complexity (e.g., using the same number of free parameters). *Comparative model fit* is an approach to theory testing that assesses the validity of competing theories by comparing their statistical fit to the data. This approach was used in many areas in behavioral science in general (e.g., Cohen, Dunbar, & McClelland, 1990) and in behavioral decision making in particular (Ganzach, 1995, 1998; Goldberg, 1970). The approach was also used to test the validity of Fast and Frugal heuristics, by comparing the fit of models based on the assumption that people rely on such heuristics in making decisions to the validity of models based on the assumption that people make compensatory decisions (Dhami, 2003; Dhami & Ayton, 2001), and the approach has been recommended by Kupec (2006).

One methodological issue that requires particular attention in this context is *cross-validation*. Certain models may achieve a better fit than others because they are closely tailored to the specific data used to set their parameters. Since one is typically not interested in the fit of a model to the particular data set used, but rather to data of the same type in general, cross-validation is the method of choice. The data must be partitioned into subsamples and model building performed based on one subsample, while the others are kept for later use in validating the initial analysis. Ideally — and this is how we will proceed — the process is repeated multiple times to increase the reliability of the estimate.

Summarizing, we will attempt to determine the complexity of the decision pattern followed by judges in routine work in a traffic court as a case study of modeling real-life decisions. We will check three aspects of their decision pattern: Are they using variables individually or do they extract higher-order interactions; do they combine them compensatorily or not; and do they exploit all the information available or use a stopping rule to decide when to ignore further information. In particular, we will try to replicate the findings by Dhami (2003), who found that a simple noncompensatory model of judges' decision-making performed better than more sophisticated and *prima facie* more plausible models.

2 Method

2.1 Data collection

We obtained permission from the President of the Court System to consult the archive files on judgments rendered by traffic judges in a large town in Israel. We obtained data from two judges in order to compare their judgments. Overall, we collected and coded data concerning 639 closed cases in the two judges (Judge A: 351 cases, Judge B: 288 cases). To gain better understanding of the ecological constraints involved in this setting, the researchers attended about 60 trials. It is held that decision makers have poor self-insight (Evans, Clibbens, Cattani, Harris, & Dennis, 2003) and we would have been interested to test this by comparing our models with the judges' own accounts of how they come to a decision. Regrettably, both declined to discuss these matters with us.

[1] Usage regarding the term "fit" is inconsistent. We will use the term throughout to refer to the proportion of cases correctly predicted by the model. Other authors use the term prediction for this, and fit to refer to the deviation between the predictions of a model and a particular data set.

2.2 Variables selection

2.2.1 Predictors

Traffic judges have wide latitude in issuing their verdicts. From their comments in court, whether addressed to the public or to the defendants, it was clear that they try to educate and to deter. They also expressed irritation when defendants did not show up in court, commented about their being already punished by having suffered bodily harm, and so forth. Based on such comments, we selected the variables they are most likely to consider, beyond the severity of the offense by itself.

We were interested in the possible interactions between severity and the other variables. This led us to concentrate on the following variables for our analysis (the name of the variable is italicized)*: Serious* (especially serious offences, such as drunken driving), *Age*, *Gender*, *Ethnicity* (Jew or Arab), number of *Previous* offences, *Vehicle* (private car, minivan, etc), *Hit* (whether bodily harm was inflicted), *Hurt* (whether the driver him/herself suffered bodily harm), *Present* (defendant present or absent), *Lawyer* (present or absent), *Experience* (number of years since obtained driver's license). Variables that have a numeric value were dichotomized at their median, except for *Age* where we dichotomized at age 21, after looking at the distribution. Many of the other variables applied only to a small number of cases, and their intersection with other such variables did not allow any meaningful analysis.

2.2.2 Punishment

Punishment consists of up to three components: fine imposition, driver's license suspension, and jail (normally offered as an alternative to paying a fine). Each component can be more or less heavy. From attendance at the trials, it was clear that both judges use these components as interchangeable to some extent. Thus, they might tell a defendant, "I ought to suspend your license, but since you need to drive for a living, I will instead . . . "

We made several ultimately unsuccessful attempts to determine the best way of weighing the punishment components (fine, prison and revocation).[2] Since we had no principled way to determine how the judges weighted the three components while determining the punishment of the offenders, we ended up using weighing the components equally for all of the components. The weights used in this study are presumably not those used by the judges, but, absent a better way of determining the latter, our approach is reasonable. Accordingly, we standardized the three components, summed them, and standardized the sums. This provided us with a numeric index of severity. For purposes of model testing we split this index at the median to create categories of punitive and nonpunitive.

3 Results

The distribution of the predictor variables was as follows: Among the offenders, there were 119 Arabs and 515 Jews; 512 Males; Average age = 34 (*SD=13*); Previous offenses = 6 (*11*); driving experience = 12 (*10*). In the incident that brought them to court, 289 drivers were themselves hurt, in 517 cases the driver hit someone else, the driver showed up on 431 occasions, and a lawyer appeared in court on 223 occasions. The files contained some additional information that we did not use, such as the defense pleas, economic status, etc.

Using the equal weights index, judges did not differ much in their overall severity. Judge A meted out on average a punishment of –0.16 and judge B of 0.19. (Recall that punishment severity was standardized with mean=0 and s.d.=1.) Across judges, licenses were suspended on average for 0.7 (*0.7*) months, a fine of 950 (*513*) shekels, that is, about USD 250 (*USD140*) was imposed, and they were condemned, at least nominally, to 8 (*11*) months in prison.

[2] Our attempts were as follows:
Expert mean judgment. Our first attempt consisted of determining the weights implicitly used by several experts, with the intention of attributing the average of these weights to the judge. Accordingly, we asked four traffic lawyers to estimate the severity of fifty-four different punishments composed of various combinations of the three punishment forms, then extracted the weights of each by multiple regression. Unfortunately, these experts turned out to weight the severity of these components so differently from one another that there was no basis to use the mean of their coefficients as an estimate of the judges' weights.
Canonical correlation. Another attempt to extract the weights was to use canonical correlation. Canonical analysis is a procedure for assessing the relationship between two sets of variables. One set was formed from variables describing the case, and the other consisted of the punishment components. The procedure calculates sets of weighted means for each set, and so defines "canonical roots" that replace the original variables and maximize the variance in one set of variables explained by the other set. Doing this, we found that the explained variance when predicting the punishment components roots from the trial data roots was low (37.2%). The different roots had diverse correlation patterns with the components of the punishments. In view of the low explained variance and the structure of correlations disclosed by the procedure, we concluded that this procedure also failed to find a good set of weights to combine the three components into a single punishment variable.
Bootstrapping. We also tried a form of bootstrapping, in an attempt to extract the relative weights the judges gave to the punishment components from their own verdicts. Our approach was to try to find pairs of components that work in tandem. For this procedure to work, one must assume that when the punishment is less punitive (e.g., women vs. men) the change in both components reflects the same change in severity (e.g., if men had a mean of 750 nis fine and 3 month prison and women had a mean of 600 nis fine and 2 month prison, we might conclude that 150 nis fine is equivalent to 1 month in prison). We worked with one pair of components at a time (e.g., prison and suspension, and prison and fine), and contrasted their values for categorical variables (such as men vs. women). However, the findings were inconsistent, depended on the categorical variable, so that, again, we could not to use these weights.

3.1 The Matching Heuristic (MH)

We will follow and expand on the approach used by Dhami (2003; Dhami & Ayton, 2001). The Matching Heuristic may be considered a variant of TTB (Take the Best). TTB in its classic form aims to force a choice between two alternatives, and tries to identify a single cue that will allow it to break the tie. The cues are dichotomous, and the decision maker looks for one of two values. Consider the decision where one must choose the more populous of two German cities. The first cue might be recognition, and the decision maker looks for recognition (rather than for failure to recognize). If one of the two cities is recognized and the other is not, this breaks the tie, and the system decides that the recognized city must be the more populous one. If it applies to neither, the system makes a decision at random. If the cue applies to both, another cue is considered. Cues are examined by some measure of validity. After all the cues have been exhausted, if none was found to discriminate, a random decision is made. The MH has the much same logic (selecting one cue at a time, with the cues ordered in advance by some measure of validity) but it applies to a single case, and its goal is to decide whether or not to make a given decision. Further, there is a default decision, such as allowing bail, or not prescribing a certain medication. Cues are processed in order of validity until one is found that indicates to take the non-default decision (not allowing bail, prescribing the medication). Failing that, the default is chosen.

Specification of the model: Variables are rank-ordered by their utilization validities. For each case, k variables are searched in order, for a critical value that indicates a punitive decision. If a critical value on a variable is found, search is terminated and a punitive decision is predicted. Otherwise, search continues until k variables have been searched, and if by this time no critical value has been found, a nonpunitive decision is predicted.

Construction of the model: The model considers variables in succession, and the sequence is determined by each variable's *utilization validity*, defined as the proportion of cases with the "critical value" that were treated punitively in the modeling set. The variables are dichotomous, and each variable has two possible values. The *critical value* for each variable is the value of that variable that was most frequently [**absolutely/ relatively**] treated punitively in the cases in the modeling set. To illustrate, consider two variables with two values each A (A_1 and A_2) and B (B_1 and B_2). Suppose the proportions of cases treated punitively are as follows: A1: 30% A2: 50% B1: 75% and B2: 70%. The critical value for A is A2 (since 50%>30%), and that for B is B1 (75% >70%). B has higher utilization validity than A, since 75%> 50%. As will be seen below, selecting the critical value advisedly has significant consequences for the number of variables that need to be consulted before a decision is made, and on the validity of that decision. The maximum number of variables the heuristic searches (i.e., *k*) is determined by systematically testing the heuristic's ability to predict correctly the decisions in the modeling set where *k* goes from 1 to n, the number of variables. The value of k that yields the greatest percentage of correct predictions is selected.

The specification of the model is not entirely satisfactory as it stands, and some preliminary work must be performed before we can engage in a meaningful comparison, which we now describe.

3.1.1 Additional specifications

Absolute vs. relative. The MH examines cues in order of their usefulness, and this has two aspects: *discrimination rate* — that is, how frequently can the cue be used to make an inference; and *utilization validity* (UV): the extent to which the cue, when used, points to the correct decision. Searching through cues by descending UV places a premium on accuracy with the potential drawback that many cues must be searched through before a discriminating cue is examined. Newell et al. (2004) found that, in a simulated stock market environment involving a series of predictions about pairs of companies, participants' pre-decisional search strategies conformed to a pattern that revealed sensitivity to both the validity and discrimination rate of cues. Rakow, Newell, Fayers, and Hersby (2005) showed that ecological validity (used to predict the environment) best predicts which cues are acquired most often. We will examine which is the best way to reflect the decisions made by the judges, whether based on discrimination or utilisational validity.

Consider a dichotomous variable A, for instance *Gender*. Combining it with *Punishment* (high/low) yields a 2 by 2 matrix. MH selects one of A's two values, called the *critical value* (e.g., *male*), and for any given case checks whether it has that value. However, which value should be considered the critical one, *male* or *female*? Assume that there are 400 males and 100 females. Assume further that out of the 400 males, 300 received a severe punishment, and out of the 100 females, 90 did. For the males, the proportion of defendants who are severely punished is 75%, for the females, it is 90%. Relatively more females are punished, then, but in absolute terms, more men than women received a severe sentence.

To understand why this matters, one must realize that every variable has potentially two different forms, and one only is used by the algorithm. When the MH considers the variable Gender, it asks one, and only one question, either: "Is it a man?" (and if so, let us be punitive, since this happened more often than the converse; if not,

we move to the next variable) or "Is it a woman?". MH does not ask both, raising the issue of which question is better. Dhami (2003) selected "the value of that variable that was most frequently treated punitively in the cases in the modeling set." The example there makes it clear that by "most frequently" is meant the value that occurs most often, in the absolute sense: if more men were treated punitively than women, then the critical value is *male*. (The experimental design in the original paper by Dhami & Ayton (2001) did not require them to choose between an absolute and a relative definition of the criterion.)

Using the absolute value ensures a frequent relevance of the rule. If there are many males, and on balance males receive a severe punishment, then using *male* as the critical value does make sense. The rule will provide guidance every time a male is encountered. This absolute version minimizes the number of comparisons made before deciding, allowing the model to function with fewer variables. Using fewer variables is often presented as an advantage, on the implicit assumption that there is a cost to a large number of comparisons. The cost in question may be in terms of number of steps and consequent duration of the process if the processing is held to be serial, as in the illustrations usually offered by partisans of the Fast and Frugal approach, or it may translate into an increased load, if the decisions are made concurrently. Several authors (e.g., Chater, Oaksford, Nakisa, & Redington, 2003; Newell, 2005) pointed out that much cognitive activity takes place in parallel and without added effort, especially with well-practiced skills, and view these considerations of costs as unconvincing (Bröder & Newell, 2008).

Alternatively, in selecting whether to test for male or for female, the model can privilege validity and use the (relative) *proportion* of cases treated punitively, and select as critical value the one with the highest proportion. If, as in our example, a higher proportion of females than of males are punished severely, then *female* is selected as the critical value to be tested: is it a female? If so, punish; if not, proceed to the next variable. The rule will be useful less often, as there are fewer females, but when it will be applied, the validity of the choice will be higher than with a rule asking about males. We will compare the two approaches directly, and will do so together with another factor to which we now turn our attention: interactive variables.

Reflecting interactions. In a laboratory study involving multi-attribute inference, Nosofsky and Bergert (2007) found that, with extended training, observers learned the relations between the attribute interactions and the criterion variable, and exploited it by recoding the interacting attributes into emergent configural cues, then applied a set of hierarchically organized rules based on the priority of the cues to make their decisions. This finding suggests that it would be best to incorporate such interactive variables before running the simulation. As we saw, each case file contained about 20 variables, detailing the defendant, the particulars of the offense charged and the defendants' responses to the charges. Before embarking on building models to test how variables are weighted, we determined whether the variables occurring in the files were best used as is. Specifically, we considered that judges might use some form of profiling, and that their decisions would be affected by combinations of traits. For instance, a recent report in the media indicated that, whereas Arabs constitute about 20% of the population in Israel, 37% of the people involved in traffic accidents that caused injuries are Arabs, and a striking 78% of casualties under the age of 19 are Arabs. According to "Green Light," a NGO fighting traffic accidents, change must come "from below" by adopting more responsible driving habits and by increased enforcement by the authorities (Stern, 28/10/2008). With this in mind, we selected various sets of variables that, based on judges' comments in court, might be involved in such interactions. The best set consisted of Age (young or adult), Ethnic (Jew or Arab), Serious (offenses of special severity) and Presence (whether the defendant appeared in person in court). For each set of variables, we performed a log-linear analysis including these four variables with the dichotomized punishment, and tested successively for the presence of any interaction of L factors, for L from 2 to 5.[3] These are simultaneous tests that all L-Factor interactions are simultaneously zero or, conversely, that there is at least one significant interaction of that level (see Table 1).

The findings concerning Judge A are presented on the left side. From the third line, one sees that there are interactions of three variables, and none of a higher order. If we think in terms of independent vs. dependent variables, this means that there are second-order interactions, and none of a higher order. We are only interested in interactions of variables that involve punishment, since these are the ones that show that cues are combined in deciding on the sentence. The ones flagged by the log-linear analysis appear in Table 2. For the cases of Judge B, there are no interactions beyond the pair-wise ones (see right side of Table 1). In terms of independent variables, only the main effects of several variables on punishment are significant. Judge A does combine variables into a higher order pattern. Certain specific profiles are singled out for special leniency or severity by Judge A, whereas we found no evidence of this approach for Judge B.

Since Judge A was sensitive to certain interactions of independent variables, the next step was to identify what

[3]Note that a five-way interaction in a log-linear model corresponds to a four-way interaction of the predictors.

Table 1: Log-linear analysis of existing order of interactions for both judges. The highest order that is still significant, indicated in bold, is the order of the interactions existing in the data set.

	Judge A			Judge B		
Order of interaction	Degrees of freedom	Max. Lik. χ^2	p	Degrees of freedom	Max. Lik. χ^2	p
2	10	111.46	0.000	**10**	**38.39**	**0.00**
3	**10**	**42.22**	**0.004**	10	10.85	0.37
4	5	7.98	0.898	5	2.98	0.65
5	1	1.96	0.903	1	0.045	0.83

specific combinations of values are singled out for leniency or severity. We considered that the combination of cue variables singled out by the judge is the one whose proportion of punishment differed most from the other three cells and is, as it were, responsible for the interaction (see Table 2, where those outliers values are in bold). For the Ethnic-Present, that value was *Absent-Arab*, for Ethnic-Born the value was *Young-Jew*.[4]

Once introduced, and in keeping with Nosofsky and Bergert (2007) we used these "interaction" variables exactly like the regular variables. Thus, with the absolute approach, we selected as critical value either the combination of values itself (e.g., Absent-Arab) or its complement (the other three cases), depending on the absolute number of cases treated punitively by each. For the relative approach, we selected the proportion of cases treated punitively. Combining the two manipulations yields four models: Hierarchical-Absolute (this is the model used in Dhami, 2003) with the critical value selected by the highest number of cases treated punitively, Hierarchical-Relative, Interaction-Absolute (interaction variables added), and Interaction-Relative.

3.1.2 Simulation.

We compared the four models by running 50 simulations of each. Each simulation consisted in selecting half the cases at random, and determining from that subset the order of the variables, as well as k, the number of variables that would be involved in the model. We then used these values to predict the punishment on the remaining cases. We calculated two indicators for each simulation. The first was the fit, that is, the proportion of cases where the punishment computed by the model matched that meted out by the judge Since the punishments were bisected at the median, the baseline is fifty percent.[5] We further recorded k, how many variables were used by the model

Table 2: Relative and absolute frequency of punishment of the interaction variables.

	ETHNIC	
Absolute number *Percentage*	Jew	Arab
Young	6	7
	14%	*70%*
Adult	*119*	37
	52%	*63%*
Present	*90*	19
	45%	*46%*
Absent	35	25
	49%	***89%***
	SERIOUS	
	Regular	Serious
Present	*88*	21
	41%	*78%*
Absent	54	6
	61%	*60%*

[4] Although log-linear analyses are robust, they rely on categorical variables. Specifically, the severity of the punishment was represented by a categorical variable and this entails data loss compared to the original continuous variable. We ran a regression analysis to check whether the interactions identified by the log-linear analysis remain significant when using the original, continuous variable. The regression analysis showed two significant interactions, Ethnic-Present ($F(1,343)=4.275$, $p=0.039$) and Gender-Born ($F(1,342)=8.191$, $p=0.004$). The Ethnic-Present interaction was significant with both procedures. Examining the data, we found our sample includes only a single female Arab driver (Bedouin women rarely drive). Gender-Born (found in the regression) and Ethnic-Born (found with the log-linear analysis) are therefore equivalent. Overall, the results from the regression analysis confirm those of the log-linear analysis.

[5] The proportion of cases classified as nonpunitive was .506. Differences in model fit therefore cannot result from differences in the proportion of "punitive" predictions.

Table 3: Utilization validity of all variables

Order	Variable Name	Utilization validity Judge A	Utilization validity Judge B
1	Absent-Arab	0.89	—
2	Serious	0.73	0.57
3	Young-Arab	0.70	—
4	Ethnic	0.64	0.53
5	Present	0.61	0.51
6	Born	0.54	0.56
7	Experience	0.53	0.55
8	Hurt	0.52	0.54
9	Hit	0.52	0.55
10	Previous offense	0.51	0.56
11	Lawyer present	0.51	0.57
12	Gender	0.51	0.52
13	Private Vehicle	0.50	0.53

for each simulation.

Both manipulations affected fit significantly. For presence of interaction variables, mean fit without interactions: 0.59; with interactions: 0.60 (SS=0.003, F(1,196)=5.3, p=0.02, SS_{Error} 0.118). For relative versus absolute, mean fit absolute: 0.57; relative: 0.61 (SS=0.082, F(1,196)=135.7, p<0.0001). The intereraction between them did not approach significance. The number of variables used (k) for each condition is shown in Table 4. Both main effects are significant: inclusion of interaction variables (F(1,196)=11.87, p<0.001) and absolute/relative (F(1,196)=187.712, p<0.001). The interaction between them was significant too (F(1, 196)=6.991, p=0.008). Summarizing these findings, more variables were used when interaction variables were introduced and when the critical values were selected by relative frequency, and both changes improved the fit.

We found that the best-fitting exemplar of Matching Heuristic uses interaction variables. Further, it uses a relative ordering of cues. These results invite the following interpretation. MH is noncompensatory. Cues are rank ordered, and the heuristic goes over the cues one by one until it finds one, C, that allows it to take a decision. That decision is final. Lesser ranking cues, that would have been consulted if C had not settled the decision, are not consulted and cannot modify the decision. To achieve a good fit, it is therefore well to use cues with high utilisation validity early on, even if their discrimination rate is not very high. This way, earlier rules apply to few cases, but when relevant and applied, they mostly yield a valid

Table 4: Mean number of variables used (k)

	Absolute	Relative
Interaction variables included	1.18	3.32
No interaction variables	1.18	2.56

decision. Cases that remain undecided by the earlier cues may still be correctly flagged by later ones. The absolute approach, by contrast, entails that the decision is often made early in the ordering by rules that have lower utilization validity. When relative frequency is used, on average, more cues are considered in coming to a decision than with the absolute version, but that decision will more often be correct.

3.2 The Weighing Model (WM)

Specification of the model. Variables are differentially weighted. For each case, variable values are multiplied by their weights and then summed. If the sum is equal to or greater than a threshold value, then a punitive decision is predicted. If not, a nonpunitive decision is predicted.

Construction of the model. Variable values are coded as 0 or 1 (for example, females were coded as 0 and males as 1 for the gender variable). A *threshold* value for predicting a punitive decision is established by averaging the sum of all such variables values over the cases in the modeling set. The *weight* for each variable is determined from the modeling set by calculating for each variable value the proportion of cases treated punitively, comparing the proportions for the different variable values, and then taking the greatest proportion as the weight for the variable.

As may be observed, the relative version for cue weighing is used for the weighing model (WM). This is indeed the only version that makes sense for this family of models, since the weighed sum it computes for each case involves always the same number of variables. The argument in favor of using the absolutes version for MH, namely that it allows to consider fewer variables, does not apply here.

Additional specifications. The specification of the two models is not yet entirely satisfactory as it stands, and we must clear two more hurdles before comparing the MH model and the WM (Weighing Model). First, Bröder (2002; Bröder & Schiffer, 2003a, 2003b) criticized how the comparison was set up by Dhami (2003; Dhami & Ayton, 2001), pointing out that MH was allowed to select the optimal number of variables it would use after comparing all the possibilities. This gave it an extra free parameter over those of the Weighing Model (WM) that always used all the variables. To ensure a fair comparison and assess the significance of this criticism, we simulated the weigh-

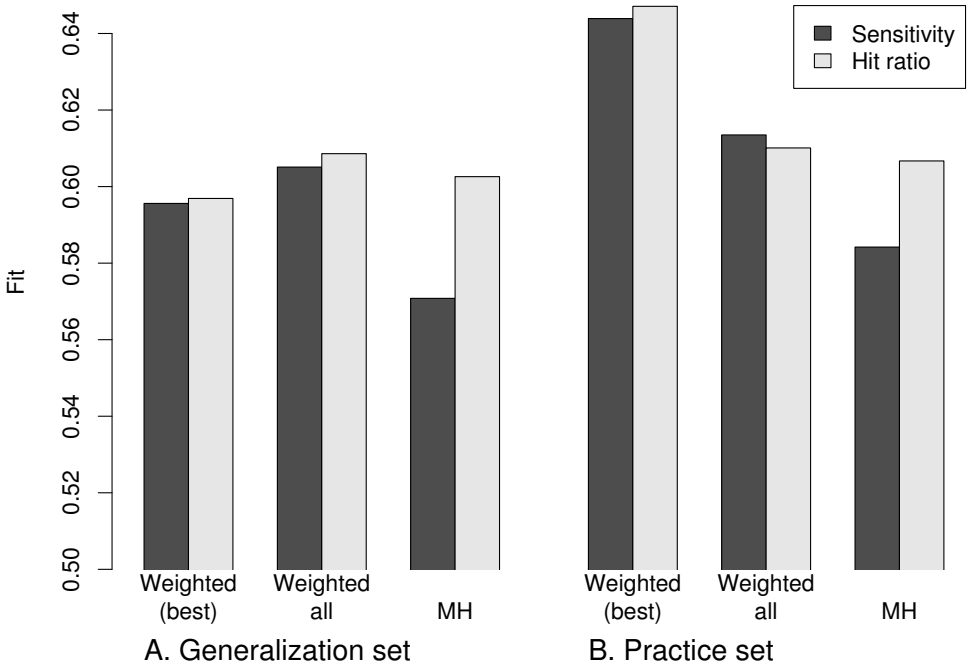

Figure 1: Fit of the six models for the generalization set, and for the practice set. Bars indicate 0.95 confidence interval.

ing algorithm both ways: using all the variables, as done by Dhami, but also allowing WM to select the best performing set of k variables. As with MH, the maximum number of variables the heuristic searches (k) is determined by systematically testing the heuristic's ability to predict correctly the decisions in the modeling set, where k goes from *1* to n, the number of variables. The value of k that yields the greatest percentage of correct predictions will be selected.

Further, the weighting of cues in WM must be examined more closely. The issue here is distinct from one we discussed and settled above, and involved selecting, for each variable, the value considered critical (i.e., should we ask about the defendant being a male, and if so making a decision, or about the defendant being a female?). We are concerned about the relative weights of the cues, for the WM model, or the order in which the variables will be considered by MH. Dhami ordered (for the MH model) or weighted (for the compensatory model) the variables by the proportion of punitive decisions for the critical value. An alternative would be to use the difference of proportions of punitive decisions between critical and noncritical cue value, a property we will call *sensitivity*. This alternative approach fits plainly the logic of the weighing algorithm, which is to give more weight to those variables that provide a more valid indication. It is less clear a priori whether this approach would also help or possibly hinder the MH model.

To illustrate, suppose the variables are A and B, and the proportion of cases treated punitively are A_1: 30% A_2: 50% B_1: 75% and B_2: 70%. The critical value for A is A_2 (since 50%>30%) and that for B is B_1 (75% >70%). The sensitivity values would be 20% for A (50%-30%) and 5% for B (75%-70%). If the variables are weighted/ordered by the proportion of punitive decisions for the critical value, for A, this would be 0.50 and for B it would be 0.75, and B would be used before A, or weighted more heavily. If sensitivity is used instead, A would weight four times as much as B, and with the MH, it would be tested first.

All told, then, we compared 3 X 2 models: (1) Weighing Model with the optimal number of variables, (2) Weighing Model with all variables included, and (3) the Matching Heuristic (with optimal number of variables). Each of these basic models was modelled for two variants: weighting/ranking the cues according to sensitivity and by the hit ratio or the critical cue value.

3.3 Comparing the two models

Having selected the optimal MH model (including interactive variables and using a relative selection of the critical value), defined the parameters of the Weighing Model (all the variables, or only the k best ones), and specified the two ways to weight/order the variables (sensitivity or hit rate frequency), we are at last ready to compare the models.

We ran the simulations as in the previous section, and performed 50 simulations and cross-validation of each model. Figure 1 shows the results. The left panel presents

the findings on the fit for the generalizations sets. Fit was defined as the proportion of cases correctly predicted by the model, and the judge's classification of each case (nonpunitive vs. punitive) was defined by a median split on the punishment measure. The interaction of Model (3) x Evaluation method (2) was significant (F(2,98)=9.2; p = 0.00023). A post hoc Neuman-Keuls test showed that the only value to depart significantly from the rest is the MH model with variables ordered by sensitivity. The hit ratio is indeed the evaluation method of choice for the MH model. The right panel of Figure 1 shows the fit values for the six models on the practice set. Comparing the two panels, it is seen that the Weighing Model with optimal number of variables is less robust than the other two.

While the differences in fit for the generalization set are mostly insignificant, the differences in k are huge. The Matching Heuristic uses less than two variables on average (1.70 when selected by hit ratio, 1.18 when selected by sensitivity) while the weighed model requires from 6.14 (sensitivity) to 8.90 (hit ratio). An ANOVA showed that both main effects of Evaluation method and of Model are hightly significant, as is their interaction F(1,49)=14.112, p=.00046. a post-hoc Neuman-Keuls test indicated that the choice of evaluation function does not significantly affect the MH, while all other values were found to differ significantly from one another (ps<.0001). As anticipated, using sensitivity rather than hit ratio to weigh and select the variables is more efficient for the Weighing Model. The same change makes a much smaller and statistically non-significant difference for the MH.

4 Discussion

We set out to analyze the complexity of the decision pattern followed by professional judges in routine sentencing in a traffic court, as a case study of modeling real-life decisions made by experienced professionals, and the adequacy of the Matching Heuristic approach to model it. Specifically, we tested the claims by Dhami (2003), who found that the Matching Heuristic is superior to the weighing Model, in a comparable setting.[6]

A preliminary question we asked was whether the judges use variables separately or extract higher-order interactions. We did find second order interactions of predictors for one judge, and indeed the largest utilization validity was for the interaction variables; the other judge did not have any use for them. On the remaining eleven cues, there was no correlation between the judges in the utilization validity (r=0.07, p=.84, see Table 3 above). This confirms the lore about these judges, as told to us informally by the lawyers. The addition of interactive variables, licensed by a suitable log-linear analysis, contributed significantly to the goodness of fit.

Comparing models such as the Matching Heuristic and the Weighing Model is more complex than might seem. When MH is used the way it was by Dhami (that is, with an absolute determination of the critical values), and compare it with the proper way to weigh variables for the WM (meaning, with variable weights determined by sensitivity), WM fares significantly better (LSD test p=.003).

To compare meaningfully the two approaches, however, we must give the models their full chance — and this means selecting the critical value by using relative rather than absolute frequency. No reason was ever offered to support the latter approach. When this is done, MH produces as good a fit as WM. Our simulations further show that the specifications of the Matching Heuristic are coherent: using the hit ratio to rank order the cues with the MH modell leads to the best fit without affecting significantly the number of variable used, whereas using sensitivity to rank order the variables impairs the fit. Similarly, the use of sensitivity to weight the cues, rather the hit ratio as done by Dhami, enables the Weighing Model to rely on fewer cues without significantly affecting fit. Finally, while allowing the WH to select the optimal number of variables (as suggested by Bröder, 2002) does not improve the fit, it does allow for fewer variables without impairing it, making for a more efficient algorithm.

Under optimal conditions, the fit of the best example of the two models was close. While the compensatory weighing model was slightly superior, the differences were not statistically significant. Further, when allowed to select the optimal number of variables to use, the Weighing Model relied on many more variables than the Matching Algorithm. When proper methodological precautions are taken, then, we did not reproduce the finding by Dhami (2003), who found the matching heuristic (MH) model superior in modeling bail granting decisions, or those of Smith and Gilhooly (2006) in modeling depression treatment. Yet it is remarkable that using only a fraction of the variables and combining them non-compensatorily, that model obtained virtually as good a fit as the Weighing Model in our simulations.

There is by now a large literature concerning the Fast and Frugal approach to decision making, in particular comparing TTB to other approaches, on the basis of laboratory work (e.g., Bröder & Newell, 2008) and analytical studies (Gigerenzer, 2008; Hogarth & Karelaia, 2007). These reviews make it clear that the appropriateness of a heuristic is very much dependent on the details of the case, including such variables as information costs, time pressure, memory retrieval, stimulus formats, or intelli-

[6]The first study by Dhami & Ayton (2001) concerned decision made by magistrates, most of whom were lay people. They too work under constraints such as time pressure, but the legal procedures, guidelines, training, expectations and responsibilities are very different.

gence. Naturalistic studies are rarer. The present study examined one case in detail, with attention to the various parameters that may affect the success of each family of models. Further research, both naturalistic and experimental, is required to generalize and qualify its findings.

References

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*, 410–433.

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 107–129.

Bröder, A. (2002). Bailing and jailing the unknown way: A critical examination of a study reported by Dhami and Ayton (2001). *Berichte aus dem Psychologischen Institut der Universität Bonn, 28* (1).

Bröder, A. (2003). Decision making with the "Adaptive Toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*, 611–625.

Bröder, A., & Newell, B. R. (2008). Challenging some common beliefs: Empirical work within the adaptive toolbox metaphor. *Judgment and Decision Making, 3*, 205–214.

Bröder, A., & Schiffer, S. (2003a). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making, 16*, 193–213.

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

Ceci, S. J., & Liker, J. K. (1986). A day at the races - a study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General, 115*, 255–266.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90*, 63–86.

Chenoweth, T., Dowling, K. L., & Louis, R. D. S. (2004). Convincing DSS users that complex models are worth the effort. *Decision Support Systems, 37*, 71–82.

Chewning, E., & Harrell, A. (1990). The effect of information overload on decision makers' cue utilization levels and decision quality in financial distress decision task. *Accounting, Organizations and Society, 15*, 527–542.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review, 97*, 332–361.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–114.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics. In G. Gigerenzer, P. M. Todd & he ABC Research Group (Eds.), *Simple heuristics that make us smart.* (pp. 97–118). New York: Oxford University Press.

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

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science, 14*, 175–180.

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making, 14*, 141–168.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115*, 199–211.

Dougherty, M. R., Thomas, R., & Franco-Watkins, A. M. (2008). Postscript: Vague heuristics revisited. *Psychological Review, 115*, 211–213.

Eppler, M., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, and related disciplines. *5*, 325 - 344.

Evans, J. S. T., Clibbens, J., Cattani, A., Harris, A., & Dennis, I. (2003). Explicit and implicit processes in multicue judgment. *Memory & Cognition, 31*, 608–618.

Ganzach, Y. (1995). Nonlinear models of clinical judgment — Meehl's data revisited. *Psychological Bulletin, 118*, 422–429.

Ganzach, Y. (1998). Nonlinear models in decision making: The diagnosis of psychosis versus neurosis from the MMPI. *Organizational Behavior and Human Decision Processes, 74*, 53–61.

Garcia-Retamero, R., Hoffrage, U., Dieckmann, A., & Ramos, M. (2007). Compound cue processing within the fast and frugal heuristics approach in nonlinearly separable environments. *Learning and Motivation, 38*, 16–34.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*, 20–29.

Gigerenzer, G., & Goldstein, D. G. (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*, 230–239.

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

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 803-+.

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

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

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

Hogarth, R. M., & Karelaia, N. (2007b). On heuristic and linear models of judgment: Mapping the demand for knowledge. *Psychological Review, 114*, 733–758.

Hutchinson, J. M. C., & Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes, 69*, 97–124.

Kupek, E. (2006). Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC Medical Research Methodology, 6*, 13–22.

Lee, B. K., & Lee, W. N. (2004). The effect of information overload on consumer choice quality in an on-line environment. *Psychology & Marketing, 21*, 159–183.

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*, 343–352.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences, 9*, 11–15.

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 999–1019.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Rakow, T., Newell, B. R., Fayers, K., & Hersby, M. (2005). Evaluating three criteria for establishing cue-search hierarchies in inferential judgment. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*, 1088–1104.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review, 109*, 599–604.

Smith, L., & Gilhooly, K. (2006). Regression versus fast and frugal models of decision-making: The case of prescribing for depression. *Applied Cognitive Psychology, 20*, 265.

Stern, Y. (28/10/2008). Green Light: 37% of those involved in fatal traffic accidents are Arabs *Haaretz,* p. 12.

Weenig, M. W. H., & Maarleveld, M. (2002). The impact of time constraint on information search strategies in complex choice tasks. *Journal of Economic Psychology, 23*, 689–702.

Youmans, R. J., & Stone, E. R. (2005). To thy own self be true: Finding the utility of cognitive information feedback. *Journal of Behavioral Decision Making, 18*, 319–341.