von Helversen, Bettina ; Rieskamp, Jörg

The mapping model: A cognitive theory of quantitative estimation

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. Journal of Experimental Psychology: General, 137(1), 73-96. https://doi.org/10.1037/0096-3445.137.1.73

Running head: MAPPING MODEL OF QUANTITATIVE ESTIMATION

The Mapping Model:

A Cognitive Theory of Quantitative Estimation

Bettina von Helversen & Jörg Rieskamp

Max Planck Institute for Human Development

Corresponding author's address:
Bettina von Helversen
Max Planck Institute for Human Development
Lentzeallee 94
14195 Berlin, Germany
Phone: +49 (0)30 8240 699
Fax: +49 (0)30 824 06 394
Email: vhelvers@mpib-berlin.mpg.de

Keywords:

decision making; multiple cue judgments; quantitative estimation, exemplar models

Abstract

How do people make quantitative estimations, such as estimating a car's selling price? Traditionally linear-regression-type models have been employed to answer this question. These models assume that people weight and integrate all information available to estimate a criterion. We propose an alternative cognitive theory for quantitative estimation: The mapping model, inspired by the work of Brown and Siegler (1993) on metrics and mappings, offers a heuristic approach to decision making. We test this model against established alternative models of estimation, namely, linear regression, an exemplar model, and a simple estimation heuristic. With four experimental studies we compare the models under different environmental conditions. The mapping model proved to be a valid model to predict people's estimates.

The Mapping Model:

A Cognitive Theory of Quantitative Estimation

Estimating unknown quantities represents a judgment problem encountered frequently in daily life. People estimate the selling price of cars, the productivity of job candidates, or the travel time for journeys. To make these estimates, people use cues that are probabilistically related to the quantity being estimated; for instance, the selling price of a car can be estimated on the basis of the car's mileage, age, or accident record. How do people make estimates? We approach this central question by introducing a new cognitive model—the mapping model. We test this model against alternative models of human estimation.

Beginning with the work of Ken Hammond (1955), who was in turn inspired by Egon Brunswik's ideas (e.g., Brunswik, 1952), linear additive models have been the standard for describing human judgments (Gigerenzer & Kurz, 2001). The research on "social judgment theory" (for an overview, see Doherty & Kurz, 1996) that followed from this seminal work encompasses a large body of studies examining people's judgments in many areas, including, among others, clinical judgments (Harries & Harries, 2001; Wryobeck & Rosenberg, 2005), teachers' evaluations of student achievement (Cooksey, Freebody, & Davidson, 1986), bail decisions (Ebbesen & Konecni, 1975), personnel selection and evaluation (Zedeck & Kafry, 1977), and medical decision making (Wigton, 1996; for reviews see Brehmer & Joyce, 1988; Brehmer, 1994). In all these studies, people's judgments are described by fitting a regression model to the data. Following the tradition of social judgment theory (Hammond, 1996) we hitherto refer to the quantity being estimated as the criterion and to the information used to estimate the criterion as the cues. Like the broader class of linear additive models, linear regression assumes that for each cue, the relation between the cue and the criterion is

abstracted and represented by a weight, where the specific weight of a cue defines the cue's impact on the final estimation.

The strong influence of linear additive models is not restricted to research on judgment and decision making. For instance, the linear additive model was employed in Anderson's (1981) "information integration theory," which describes integration of social as well as physical information. Likewise it was adopted to describe the impact of social norms on behavior (Fishbein & Ajzen, 1980). Despite the model's success in describing human behavior, in the present article we challenge the assumption that the underlying cognitive process of human judgment follows the additive integration of weighted information. In its stead we propose the mapping model as a new model of human estimation. This model is based on Brown and Siegler's (1993) work on metrics and mapping. Our main goal was to test this model rigorously against a linear additive model, and additionally against alternative recent cognitive models of human estimation.

The Mapping Model

Brown and Siegler (1993; see also Brown, 2002) suggested that real-world quantitative estimations rely on knowledge about the *mapping* properties of the objects and the *metric* properties of the criterion. The mapping properties reflect the ordinal relations among the objects in one domain, that is, the knowledge about which object will have a higher value on the criterion compared to other objects. Knowledge about the metric properties, on the other hand, refers to the statistical properties of the criterion, such as the mean, the median, and the functional form of the distribution. Brown and Siegler (1993) assumed that to make accurate quantitative estimations, knowledge about both types of properties is indispensable, yet they did not specify a computational model describing human estimation. Therefore we suggest one that is inspired by the ideas of mapping and metrics.

The mapping model specifies how knowledge about the mapping and metric properties of objects is acquired in two separate steps. First, knowledge about the mapping properties is gathered from the cues. The sum of the cue values is used to infer the ordinal relations of the objects and to group them into categories. Second, to represent the metric properties of the criterion, a typical criterion value is derived for each category by considering the criterion values of other objects falling into the same category. The mapping model only uses binary cue information, so that each cue can have either a positive or a negative value. Cues are coded so that they are positively correlated with the criterion. The knowledge about the mapping properties is then derived by a simple counting strategy, adding up the positive cue values for all cues *J* of each object *i* and categorizing them according to their cue sums:

$$k_i = \sum_{j=1}^{J} c_{ji} \qquad (1)$$

where *k* denotes the cue sum of object *i* and c_{ji} refers to the cue value of object *i* on cue *j*.

For each cue sum category a typical criterion value is abstracted, represented by the median criterion value of all known objects that share the same cue sum.[1] To estimate the criterion value of a new object, the probe (*p*), the cue sum of the probe is computed and the typical criterion value of the corresponding cue sum category is used as an estimate:

$$\hat{y}_p = Mdn\left(x_i, k_i = k_p\right), \qquad (2)$$

where $\hat{y}_p$ denotes the estimated criterion value for probe *p*, which is estimated by the median (*Mdn*) of the criterion values of all known objects *i* that belong to the group of objects with the same cue sum *k* as the probe *p*. If a cue sum category does not exist because no object with a corresponding cue sum was encountered in the past, the average value of the adjacent categories is employed as an estimate.

We demonstrate the mechanism of the mapping model with the illustrative example of estimating the selling price (i.e., the criterion) of two mobile phones, let's call them Psi and Omega, offered in an online marketplace. The phones' features (i.e., weight, display size, digital camera, and Internet access) can be employed as cues to estimate the selling price. To estimate the selling prices of Psi and Omega we can compare them on the features to four similar phones, A, B, C, and D, that were sold in the past (see Table 1). The mapping model estimates that phone Psi will sell for $100, because of the four phones sold (A–D), only phone D—which sold for $100—falls into the same cue sum category. For phone Omega with a cue sum of one, the mapping model estimates the median price of the two phones A and B with the same cue sum, which sold for $10 and $20, respectively, yielding an estimated selling price of $15.

Alternative Theories of Estimation

With the mapping model, we question the widespread assumption in cognitive psychology that human judgments follow a linear additive process of information integration. We first test the mapping model against the most established representative of linear additive models—linear regression. Because other models have recently been proposed to explain estimations from multiple cues, the mapping model is also tested against two of these competitors: an exemplar model (Juslin, Olsson, & Olsson, 2003b) and a heuristic strategy (Hertwig, Hoffrage, & Martignon, 1999). We use our illustrative example to explain the models and show how their predictions differ.

Multiple linear regression. Linear additive models assume that explicit cue–criterion relationships are abstracted and represented as cue weights. Multiple linear regression (MLR) computes optimal weights for every cue, minimizing the squared deviations of the prediction from the criterion (e.g., Cohen & Cohen, 2003). The weights indicate how much impact a

given cue has on the estimate of the criterion. The estimated criterion value, $\hat{y}_p$, of the probe p is given by the sum of the product of the cue values, c_j, of the cues j with their respective weights, ω_j, plus an intercept, ω_0:

$$\hat{y}_p = \sum_{j=1}^{J} \omega_j c_j + \omega_0 \quad (3)$$

In our example, the four sold phones are used to fit the regression model. That is, the model finds the weights that minimize the squared deviation of the predicted from the real criterion value of the phones sold. In our example optimal weights for the cues are 80, 10, 10, and 0, respectively, with an intercept of 10. The fitted regression model then predicts a selling price of \$110 and \$90 for the new phones Psi and Omega, respectively.

In addition we tested two simplified versions of this standard regression model. First, we included a stepwise regression model that includes only significant parameters (Hastie, Tibshirani, & Friedman, 2001). Second, we tested a simplified version of the regression model that was not fit to participants' estimations. Instead, the optimal parameters for solving the task were selected a priori based on the objective criterion values. However, across all of the following studies the standard regression model was most successful in predicting participants' estimations for new independent observations that were not used to estimate the models' parameters, so that for the sake of clarity we only report the results for the standard regression model.

Exemplar-based model. A promising alternative approach to quantitative estimation is provided by exemplar-based models (EBMs), which in the past have been successfully applied to explain human categorization (for an overview see, for example, Nosofsky & Johansen, 2000). Exemplar models assume that people categorize objects by determining how similar they are to formerly encountered exemplars of the categories and assigning them to

the category with the most similar exemplars. Thus, in contrast to a linear additive model, exemplar models do not assume the abstraction of cue–criterion relationships but rely on a knowledge base of memorized exemplars. Recently, Juslin et al. (2003b; Juslin, Jones, Olsson, & Winman, 2003a) reformulated the original context model of Medin and Schaffer (1978) for the area of quantitative estimation (see also Dougherty, Gettys, & Ogden, 1999; Juslin & Persson, 2002; Smith & Zárate, 1992). Juslin, Karlsson, and Olsson (in press, see also Olsson, Enqvist, & Juslin, 2006) showed that exemplar models are more suitable for predicting people's estimations than linear regression when the cues are non-linearly related to the criterion.

The exemplar model proposed by Juslin et al. (2003a, b) is closely related to the generalized context model[2] (Nosofsky, 1986, 1992; Nosofsky & Johansen, 2000). Exemplar models assume a memory-based inference process. To estimate the criterion of a new object (the probe), the similarity of the probe to the exemplars retrieved from memory is determined. The more similar the probe is to an exemplar, the closer the estimate will be to the exemplar's criterion value. The final estimate of the criterion is the average of the criterion values of all memorized exemplars, weighted by their similarities to the probe:

$$\hat{y}_p = \frac{\sum_{i=1}^{I} S(p,i) \cdot x_i}{\sum_{i=1}^{I} S(p,i)}, \qquad (4)$$

where $\hat{y}_p$ is the estimated criterion value for probe p; S is the similarity of the probe p to the stored exemplars i with the criterion value x_i; and I is the number of stored exemplars in memory. The similarity S between the probe and an exemplar is determined by the multiplicative similarity rule of the context model (cf., Medin & Schaffer, 1978):

$$S(p,i)=\prod_{j=1}^{J} d_j \,, \qquad (5)$$

where the variable *d* specifies the similarity between the probe and the exemplar on the cue dimension *j*, and d_j takes the value 1 if the values of the probe and the exemplar on cue dimension *j* match and s_j if they do not. The parameter s_j is an attention weight parameter capturing a cue's importance for the similarity assessment and varies between 0 and 1. A large value for the attention parameter *s* close to 1 implies that a mismatch on this cue has almost no effect on the overall similarity, whereas a low value for *s* close to 0 implies that the cue is very important, because the overall similarity approaches zero if the cue values do not match.

The standard exemplar model assumes that the importance given to each cue varies by using different attention parameters (e.g., Juslin et al. 2003a, b). However, by having one free parameter for each cue the exemplar model is relatively complex and it is an open question whether this complexity is required to capture the underlying cognitive process of estimations. To answer this question we additionally implemented a simplified version of the exemplar model, which assumes that only one single attention parameter *s* is used for all cues (see also Juslin & Persson, 2002). This single parameter then represents the gradient of the similarity function; that is, if *s* is close to 0 only very similar exemplars will influence the estimation, but if *s* is close to 1 also less similar exemplars will be considered. Finally, we implemented a third version of the exemplar model that did not fit parameters to participants' estimations; instead, the parameter values were derived by using the objective criterion values of the objects in the training phase. It turned out that the simplified exemplar model with only one free parameter was most successful in predicting individuals' estimations for new independent observations, so that for the sake of clarity we only report the results for the

simplified exemplar model with the exception of the following simulation study and Study 4. When the simplified exemplar model is applied to our phone example, using an attention parameter of s = .001 to predict the phones' selling prices, the selling prices of phone Psi and Omega were estimated to be $20 and $43, respectively.

A heuristic for estimation—QuickEst. Although regressions models are able to describe the outcome of a cognitive process (i.e., the final estimation), they have been criticized for not capturing the process itself (Brehmer, 1994; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Hoffman, 1960; for a review see Doherty & Brehmer, 1997). Gigerenzer, Todd, and the ABC Research Group (1999) have argued that the cognitive process of making judgments can often be best described with simple heuristics. Recent experimental work has illustrated that simple heuristics can predict people's inferential choices well, in particular when the application of complex strategies is more costly (e.g., Bröder, 2000; Bröder & Schiffer, 2003; Rieskamp, 2006; Rieskamp & Hoffrage, 1999; Rieskamp & Otto, 2006). In this vein, Hertwig et al. (1999) proposed a heuristic for quantitative estimations, QuickEst, that uses only a small amount of information. According to the heuristic, people process cues sequentially and stop searching as soon as a cue has a negative cue value. Hertwig et al. showed that QuickEst's predictions are as accurate as those of linear regression when applied in an environment where the distribution of the objects' criterion values is J-shaped. A distribution is called J-shaped if most values are small and only a few high values exist, such as, for instance, the distribution of incomes.

QuickEst uses only binary cue information. Each cue can have either a positive or a negative value. All cues are coded such that they correlate positively with the criterion. Accordingly, for each cue, objects with a positive cue value will on average have higher criterion values than objects with a negative value. Next, for each cue the mean criterion

value of all objects that have a negative cue value is computed, here called the *nil mean size* (Hertwig, Hoffrage, & Sparr, 2007). Likewise the mean criterion values of the objects with a positive cue value are determined (conditional positive mean). The idea of QuickEst is to stop searching for more information as soon as it becomes probable that an object has a small criterion value. Thus QuickEst stops search as soon as a cue with a negative cue value is encountered or if the cue value for the object is missing. If a positive cue value is encountered, the next cue is considered until all relevant cues have been looked up. QuickEst searches through the cues according to their nil mean size beginning with the smallest.

In contrast to Hertwig et al. (1999) we assume that the maximum number of cues that are searched for is a free parameter capturing individual differences. An estimation is based on the cue that stopped search. If the search was stopped because a negative cue value was encountered, the nil mean size of that cue is used as an estimate. If search was stopped because the maximum number of cues had been considered, the conditional positive mean of the last cue is estimated. For the estimates the means are rounded to the next spontaneous number[3] (Albers, 2001). For our phone example, the nil mean sizes of the cues are 20, 55, 15, and 25, respectively. QuickEst starts search by looking up the information of the phones' weight, the cue with the smallest nil mean size. If this cue has a positive value, it continues search, considering whether the phone has a digital camera, and so on. Because phone Psi has a positive value on weight and has a digital camera, search continues until information for display size is looked up. As phone Psi has a negative value on display size, the rounded conditional mean of this cue ($30) is estimated as the selling price. For phone Omega, search stops after looking up information for weight, and its nil mean size of $15 is estimated as the selling price.

Testing the theories. Conceptually the theories we consider can be distinguished by two aspects: (1) the way they abstract knowledge from objects encountered in the past, that is, their *knowledge abstraction assumptions*; and (2) the way the abstracted knowledge and the information a probe provides is processed to make a final estimation, that is, their *process assumptions*.

The regression model assumes an additive estimation rule. To build this estimation rule it abstracts knowledge about the cue weights from the encountered objects, taking the dependencies between cues into account. Once this rule is established, previously encountered objects can be forgotten. For the estimation process the model integrates all available information, determining a weighted sum of the cue values. Like the regression model, QuickEst assumes that knowledge, that is, the mean criterion values of the cues, is abstracted from encountered objects. However, QuickEst does not integrate any information; instead cues are searched sequentially and an estimation is made on the basis of one single cue. The exemplar model does not abstract much knowledge; instead it assumes that all encountered objects are stored in memory. Nevertheless, the knowledge of how much attention a cue receives is abstracted from the encountered objects. For the estimation process the exemplar model assumes that the information of all stored exemplars is integrated, by determining a mean of the retrieved criterion values weighted by the similarity of the retrieved exemplars to the probe. In sum, the regression model assumes heavy knowledge abstraction from encountered objects and an information integration process for estimation. QuickEst assumes knowledge abstraction and no information integration, and the exemplar model assumes little knowledge abstraction but relies heavily on integration of information for making an estimation.

Similar to QuickEst and regression, the mapping model assumes a rule-based estimation process, relying on the abstraction of knowledge. The mapping model groups objects into categories on the basis of their cue sums, regardless of the pattern of cue values. For each cue sum category the criterion values of the objects falling into this category are stored (see also Footnote 1). For the estimation process the cues' information on the probe is integrated by a simple adding strategy. Then for each probe the median criterion value of the corresponding cue sum category is retrieved and used as an estimate.

How does the mapping model compare to the other models? The mapping model resembles QuickEst in the way it abstracts knowledge by categorizing objects into groups. However, while QuickEst bases its estimation on only one cue, the mapping model assumes that the available information is integrated. Similar to regression, the mapping model relies on an additive integration of information. However, it assumes that every cue contributes equally to the cue sum, whereas the regression model assumes differential weighting of cues. Further, the estimation process of the mapping model does not terminate with the integration of the cue values but continues with determining the typical criterion value using the median of the criterion values of the objects falling in the same cue sum category. As in the exemplar model, this retrieval process of the mapping model can be conceptualized as "similarity based," because the retrieval is guided by finding the best match between the cue sum category determined for the probe and the criterion values for the categories abstracted from the objects encountered in the past. However, the exemplar model and the mapping model differ in how they define similarity. The exemplar model assumes that objects are represented in terms of discrete cue values and similarity is a function of the matches and mismatches on each cue. For the mapping model similarity is a strict function of the cue sum category. Thus, although the simplified exemplar model and the mapping model both assume that cues are

equally weighted, two objects that the mapping model groups together because they share the same cue sum could be very different for the exemplar model depending on the pattern of cue values.

Although the theories that we consider differ conceptually, empirically they often lead to similar predictions. To test the theories against one another it is therefore important to identify conditions under which the predictions differ. One aspect of the environment has already been shown to differentiate the theories: the distribution of the criterion values. Hertwig et al. (2007) found that QuickEst outperformed linear regression if the criterion distribution was J-shaped but performed poorly when the criterion was uniformly distributed. In J-shaped distributions characteristically only a few objects have high criterion values, while most have low values. Such distributions are so named because they resemble a J (rotated 90 degrees clockwise) if the objects are ordered according to their ranks. Formally, they can often be described by a power function (i.e., $y = b \times x^a$). A distribution following a power law additionally implies that the rank of an object is specifically related to its size, so that if log rank is drawn against log size, a straight line results. Likewise we will refer to a uniform distribution as a linear distribution, because a straight line results if rank is plotted against size.

The use of a criterion that follows a power function has a further advantage. Test situations that allow discrimination between models often consist of highly artificial cases that are no longer representative of the original problem. Power law distributions, on the other hand, are among the most prevalent distributions encountered in everyday life. Since power law distributions are related to general growth processes (Gabaix, 1999), they can well describe phenomena as diverse as people's incomes, magnitudes of earthquakes, sales of books or music, or the sizes of computer files, moon craters, or cities (Levy & Solomon,

1997; for a review see Schroeder, 1991). Therefore we extend the work of Hertwig et al. (2007) by conducting a simulation study to investigate how all the models discussed here, especially the mapping model, perform in an environment with a J-shaped and a linearly distributed criterion, respectively.

Simulation study

The goal of the simulation study was to examine how accurate the various models are in solving estimation problems under different environmental conditions. Furthermore the goal was to identify environments in which the models make distinct prediction that allow an experimental test.

The simulations were designed to resemble an experimental condition as closely as possible, while still providing enough data to result in reliable evaluations of the models' accuracies. First, J-shaped and linearly distributed criterion values, ranging between 2 and 100, were created for 50 objects by using a power function ($y = bx^a$, with $a = -1$, $b = 100$, and x ranging between 1 and 50) for the J-shaped environment and a linear function ($y = bx + c$, with $b = -2$ and $c = 102$) for the linear environment. To investigate if potential accuracy differences would hold over a wide range of situations, we varied two further factors: The cue–criterion correlation and the percentage of positive and negative cue values per cue (for details see Appendix A).

We examined models' accuracies by cross-validation (averaged over 100 trials). That is, we randomly selected 100 times one half of the data—the calibration sample— to estimate the models' parameters, and then we tested the models on the other half—the validation sample—to test the models' accuracies for new objects. Models' accuracies were characterized by the root mean square error (*RMSE*) of the models' predictions and the criterion values.

How accurate were the four models? In general, accuracy was strongly affected by the distribution of the criterion value. In the linear environment, the *RMSE* was on average two times larger than in the J-shaped environment. When fitting the data of the calibration sample, all four models performed better than a baseline model, which always predicted the average criterion value of all objects of the calibration sample (see Table 2). The exemplar model was the best model in both conditions, and QuickEst was worst. However, the validation sample represents the crucial situation of making predictions for new objects. Here QuickEst performed best in the J-shaped environment, and the mapping model was second best, $t(31) = 2.15$, $p = .02$, with an effect size of $d = .45$ (Cohen, 1988). In the linear environment, the mapping model was the best in the validation sample, followed by the exemplar model, $t(53) = 3.08$, $p < .01$, $d = .42$, and QuickEst performed worst. These results illustrate that the criterion distribution influences models' accuracies differentially. They are in line with the results of Hertwig et al. (2007), who reported that the accuracy of linear regression is affected negatively by a skewed distribution, whereas the accuracy of QuickEst deteriorates if the criterion is linearly distributed.

The difference in model accuracy between the calibration sample and the validation sample highlights the problem of over-fitting: Complex models with several free parameters are highly flexible in fitting any data, running the risk of fitting noise instead of fitting systematic structure (see Olsson, Wennerholm, & Lyxzén, 2004; Pitt, Myung, & Zhang, 2002). For this reason in our experimental studies we tested the models by using a generalization test (cf., Busemeyer & Wang, 2000): First, participants made estimations for a training set, which was later used to estimate the models' parameters. Then they made estimations in a test set, which was used to test the models' predictions against each other.

Study 1

Study 1 was designed to test how well the four models of quantitative estimation can predict human estimations. To control for prior knowledge, participants were presented with an artificial inference problem. Following the work of Juslin et al. (2003a, b), participants had to estimate the toxicity of fictional bugs, which were described by five dichotomous cues. For a rigorous test of the models, the experiment varied the distribution of the criterion values in a between-subjects design. In the first condition, the linear environment, the criterion values were linearly distributed, whereas in the second condition, the J-shaped environment, the distribution of the criterion values followed a power law function.

Method

Participants. Sixty participants took part in the experiment: 30 women and 30 men. The participants were randomly assigned to the two experimental conditions, balanced for gender. They were on average 25 years old and most were students from one of the Berlin universities. The data of one participant in the linear environment was later excluded because the participant did not put any effort into solving the task, responding with the same number as an estimate in every trial. Participants were paid according to their performance in the task; the average payment was €13 for an individual session lasting on average 1.5 hr (with €1 corresponding to $1.28 at the time of the study).

Procedure and materials. The study was conducted as a computer-based experiment. Written instructions informed the participants that their task was to estimate the toxicity of different bugs on the basis of five binary cues (color of head, length of antennae, color of wings, size, and biotope). The toxicity of the bugs was measured by the amount of venom in the saliva and could vary between 20 and 1,000 mg per liter. As a cover story the participants were told that the toxicity of the bugs differed depending on the subspecies the bugs belonged

to and that the cues would help them to estimate the bugs' toxicity correctly. The bugs could not be distinguished solely on the basis of the cues, as some of the subspecies were very similar in appearance. In these cases only a genetic test could identify the correct subspecies. To speed up learning of the task, the participants were informed about the direction of the cues, that is, which cue values indicated higher levels of toxicity, without learning the magnitude of the correlation.

Depending on the experimental condition the criterion was either J-shaped or linearly distributed. In both conditions, the experiment consisted of a learning phase, in which the participants could learn to estimate the bugs' toxicity, and a test phase, in which the toxicity of new bugs had to be estimated. In the training phase the participants had to estimate the toxicity of 20 bugs. This phase consisted of 200 trials structured in 10 blocks, each presenting the 20 bugs from the training set in random order. The participants were not told that the same bugs would be repeated; instead each time a bug reappeared, it had a new number. In each trial one bug was presented with its five cue values on the screen and participants were asked to give an estimate of the toxicity of the bug. The order in which the cues were presented was randomly determined for each participant.

After making the estimation, participants were given feedback about the accuracy of their estimate and received points accordingly. Participants' payment was contingent on their performance. After the experiment the total number of earned points was exchanged into euros at a rate of €0.1 for 100 points. For each estimation that exactly matched the correct criterion value, the participants were awarded 100 points. Deviations from the correct criterion value led to fewer points, with increasing inaccuracy leading to a disproportionately larger decrease in points. Specifically, the feedback algorithm used the mean squared

deviation of the estimation from the actual criterion value to determine how many points were subtracted from the maximum 100 points for an exact estimation.

To create a moderately exacting feedback environment (Hogarth, Gibbs, McKenzie, & Marquis, 1991), which has been shown to lead to high performance (Gonzalez-Vallejo & Bonham, in press), the feedback algorithm incorporated a correction term to account for the difficulty of the task (see Appendix B for details). The correction term consisted of a constant that determined the magnitude of the deviation that would result in a payoff of zero points. Any deviation exceeding the deviation by the correction term would lead to the subtraction of points. The correction term was chosen so that reliance on a baseline model that always estimated the same value would result in zero points. Since the baseline model reached a better fit in the J-shaped environment, the correction terms in the two environments differed. In both conditions participants received 100 points for a correct answer; in the J-shaped environment a maximum of 355 points was subtracted for an error whereas a maximum of only 127 points was subtracted in the linear environment. In the instructions it was explained to the participants that subtracting points for errors was employed to correct for chance performance.

In addition to earning points, the participants received outcome feedback on each bug's actual criterion value, the mean squared error of their estimation, and their current total score. In the test phase the participants made the same judgments as in the training phase, but without outcome feedback. They were informed that nevertheless they would earn points according to their accuracy. The test set consisted of 21 profiles that included the old profiles from the training set as well as new profiles.

The training and the test set were constructed so that the models' predictions for the test set, given the training set, would be as different as possible.[4] To find a training set–test

set combination that would allow for good discrimination between the models in both environments, we first chose an environment from the simulation in which the models had differed in their predictive accuracy. In this environment each cue had 50% negative cue values and correlated positively with both criteria. We randomly selected 100 training sets of 20 bugs from this environment under the constraint that the highest and the lowest criterion value were always included, ensuring the full range of the criterion for the estimations. All criterion values were multiplied by 10 to have a larger range. Then each model was fitted to the bugs of the training sets, maximizing the model's accuracy in estimating the bugs' toxicity. After fitting the models' parameters, the models' predictions were determined for all objects that did not appear in the training set.

From the 100 training sets we selected the one that allowed the best discrimination between all four models on the new objects, given two additional restrictions. First, to avoid the objection that the participants simply learned to make estimations according to the best performing model in the training set, we excluded all training sets in which the models' accuracy differed widely in the J-shaped environment. Second, we excluded all training sets in which the same cue profile appeared more than four times, to ensure that the differences in model predictions were not due to an extreme training set. Finally from the remaining training sets the one that maximized the differentiability of the models in the test set was selected, which was the set with the highest number of cue profiles for which two models made predictions differing by more than 100 mg/l of estimated toxicity.

The final training set consisted of 20 objects with 20 different criterion values, but with only eight different cue profiles, so that one profile appeared once, three profiles twice, three profiles three times, and one profile four times (see Table 3). All cues correlated positively with the criterion and the cue–criterion correlations differed between .30 and .79

(see Table 4). For the test set, the cue profiles for which the four models made the most different predictions were selected. For any pairwise model comparison, at least four profiles allowed a good differentiation between the two models. Also the bugs of the training set were included in the test set. The test sets of the linear and the J-shaped environments and the models' predictions based on the training set can be found in Appendix C.

How well could the different models solve the estimation problem in the training phase? In the J-shaped environment the models' predictions, when fitting the parameters to the objective criterion values, deviated from the criterion with a mean root mean square deviation (*RMSD*) of 136 and could explain about 64% of the variance. QuickEst and the mapping model did slightly worse than the other models. Because the training set in both environments consisted of the same cue profiles, the models' accuracies could not be controlled for in the linear environment, but the accuracies did not differ substantially among the regression, exemplar, and mapping model (M = 145, SD = 11). Only QuickEst with an *RMSD* of 183 did clearly worse than the other models. Although the *RMSD* in the linear environment was higher on average, the models could explain more linear variance (average r^2 = .74).

Results

Overall, the mapping model explained the predictions of the participants best in the test phase, if all conditions were considered jointly. However, the distribution of the criterion played an important role. In the J-shaped environment the mapping model was clearly the best model, whereas in the linear environment the standard regression model and the exemplar model with only one parameter performed equally well. Before we come to the model comparisons, we first report participants' accuracy.

Accuracy of the participants. Participants' accuracy was measured by the *RMSD* between participants' estimations and the criterion and by the Pearson correlation of the estimations with the criterion. Participants were quite successful in learning the bugs' toxicity levels during the training phase, in particular when considering that due to the indistinguishable cue profiles perfection was not possible. The strongest learning effects were observed between the first and the fourth block. Overall, the mean *RMSD* dropped in both environments from 236 (J-shaped) and 232 (linear) in the first block to 149 and 194, respectively, in the 10th block. The last three blocks showed no significant learning effects, so the data were merged for the further analyses. The average accuracy in the linear environment (*RMSD* = 210) was worse than in the J-shaped environment (*RMSD* = 163), U = 104, $p < .01$. However, the average amount of variance explained did not differ; $r^2_{\text{linear}} = .58$, $r^2_{\text{J-shaped}} = .58$.

Estimating the models' parameters. As the primary measure of the models' goodness-of-fit, the *RMSD* between the participants' estimations and the models' predictions was used. The models' parameters were estimated by minimizing the *RMSD* for participants' estimations in the last three blocks of the training phase. The models were tested against each other on the basis of the *RMSDs* of their estimations for the test phase. Additionally we considered the degree of linear variance explained by the models (the coefficient of determination r^2), because the two measures capture slightly different aspects of model fit and r^2 is the preferred measure in the social judgment theory literature. But since the two measures are not independent all model tests are solely based on the models' *RMSD*.[5]

The models were fitted individually to each participant: For the linear regression the parameters were determined analytically using the cues of the training set and the individual participants' estimates. The exemplar model was fitted on the last three blocks of the training

phase with the correct cue and criterion values of the training set as the memory base. The best parameter for each participant was searched for by using the quasi-Newton optimization method as implemented in MATLAB. To avoid local minima, parameters were first derived by a grid search with the results serving as the starting values for the subsequent fitting procedure. For QuickEst only one parameter had to be estimated specifying the maximum number of cues considered, and here the optimal parameter value was selected by an exhaustive search. If different numbers of cues reached the same fit the lowest number was selected. The mapping model entails no free parameters, so no parameter was estimated; the medians for the different categories the mapping model used were determined on the basis of the objects' criterion values in the training set.

Model comparison—training phase. We first compared each model's fit with the fit of a baseline model in the training phase, which predicted only one single value for all objects encountered; the specific value the baseline model predicted was fitted to the data of the training phase. The baseline model reached an average fit of *RMSD* = 289 in the linear environment and of 225 in the J-shaped environment. Because the baseline model is a rather naïve model of estimation, any of our four models needs to prove first that it can do better by taking the dependencies of the estimations on the cue profiles into account. For the training phase all four models did better than the baseline model in predicting participants' estimations (see Table 5). To test if one model could explain participants' estimations significantly better than another model we used a non-parametric test (i.e., the Wilcoxon *Z*-test). In describing the data of the training phase, the regression model did best in both environments, followed by the exemplar model (linear: MLR vs. EBM, $Z = -4.67$, $p < .01$; J-shaped: $Z = -4.78$, $p < .01$), explaining more than 80% of the variance in the linear environment and almost 80% in the J-shaped environment (see Table 5). The mapping model

and QuickEst did significantly worse than the other two models, particularly in the linear environment. As described above, for clarity we only report the results for the standard regression model with six free parameters and the simplified exemplar model with one free parameter (for the results of the other versions see Appendix D).

However, the models' fit for the training phase is not very meaningful for testing the models against each other: Even though we tried to put the models on more equal footing, they still differed in their complexities, that is, in the number of free parameters and the complexity of their functional form. Thus it is not surprising that the models with greater flexibility—the regression and the exemplar model—did better in fitting the data than the mapping model. Therefore the crucial model comparison test consists of how well the models predict participants' estimations for new independent objects of the test phase. This generalization test goes beyond a pure cross-validation test, because the new objects of the test phase differed from the objects of the training phase.

Model comparison—test phase. The models' predictions for the test phase were determined on the basis of the estimated parameters of the training phase. The baseline model reached a better fit in the test set than in the training set with an average fit of $RMSD = 180$ in the J-shaped environment and $RMSD = 282$ in the linear environment. This is presumably because the new profiles included in the test set had less extreme cue profiles; that is, the new profiles had a maximum of only four positive cues and a minimum of one positive cue (see Appendix C). In the linear environment, the regression model, the exemplar model, and the mapping model did better, on average, than the baseline model (baseline vs. EBM: $Z = -4.68$, $p < .01$). In the J-shaped environment, QuickEst and the mapping model were able to beat the baseline model (QuickEst vs. baseline: $Z = -2.05$, $p = .04$), while the exemplar model could not be distinguished from the baseline model (EBM vs. baseline: $Z = -1.37$, $p = .18$) and the

regression model performed worse than the baseline model (MLR vs. baseline: Z = −3.47, $p <$ .01).

Figure 1 illustrates the models' different successes in predicting participants' estimations. The figure shows the models' and participants' average estimations for each profile of the test phase, demonstrating that in the linear environment it is difficult to discriminate between the models, whereas in the J-shaped environment the mapping model predicted participants' estimations best. In the linear environment the regression model, the exemplar model, and the mapping model performed equally well and significantly better than QuickEst (QuickEst vs. MLR: $Z = -4.5$, $p < .01$; see also Table 5). In the J-shaped environment the mapping model was the best model in predicting the estimations (mapping model vs. EBM: $Z = -3.2$, $p < .01$) and the exemplar model was indistinguishable from QuickEst (QuickEst vs. EBM: $Z = -.03$, $p = .98$), but both the exemplar model and QuickEst outperformed the regression model (QuickEst vs. MLR, $Z = -3.59$, $p < .01$).

To consider individual differences, we examined which model, including the baseline model, was best in predicting each participant's estimations (according to the *RMSD*). In the linear environment, the mapping model was best in predicting the estimations for 12 participants (41%), the regression model was best for 11 (38%), the exemplar model for 5 (17%), and QuickEst for 1 (3%) participant. In the J-shaped environment, the mapping model was best for 16 participants (53%), QuickEst for 6 (20%), and the baseline model for 2 (7%). The regression model and the exemplar model, respectively, predicted the estimations of 3 (10%) participants best. In sum, the individual analyses led to the same conclusions as the analysis of the aggregated results: The mapping model was the best model in predicting participants' estimations. It did as well as the regression model for the linear environment, and it was the outstanding model for the J-shaped environment.

Discussion of Study 1

Study 1 showed that the mapping model was able to predict participants' estimations well in both environments, suggesting that it could be a simple alternative to standard estimation models. Although in the linear environment all models performed equally well, the exemplar model and the regression model made worse predictions compared to the mapping model in the J-shaped environment. Even though Juslin et al. (2003a, in press) showed that the exemplar model performed well in a related task, in our study, people apparently did not rely on an exemplar-based estimation process. However, this conclusion needs to be limited to the experimental situation considered, which might have been disadvantageous for an exemplar-based process. In particular, some of the cue profiles in the experiment were indistinguishable. Although this is a realistic feature in quantitative estimations in everyday life it could nevertheless have impeded an exemplar-based inference process, by making it more difficult to establish memory traces for the exemplars. Therefore in Study 2 an experimental situation was created that should favor an exemplar-based inference process and should increase the differentiability of the models in the linear environment.

Study 2

The first goal of Study 2 was to examine the reasons for the poor performance of the exemplar model in Study 1. As described above, using objects with identical cue profiles but with different criterion values could have made memorization of exemplars cognitively very demanding. Therefore in Study 2 each cue profile appeared only once in the training set. Additionally, the objects (i.e., bugs) were given names to emphasize that the same objects had to be evaluated several times. This procedure made memorization of exemplars easier and fostered an exemplar-based inference process. It also allowed, in principle, perfect

performance in the training phase, when following an exemplar-based estimation process. Thus, Study 2 provided good conditions for the exemplar model.

To test the exemplar model against the mapping model, in the test phase of the experiment all possible cue profiles that could be created with the limited number of cues were presented to the participants, so that some of the profiles had been encountered before in the training phase and some were new. The objects of the training set were presented with new names in the test phase to test for memory effects of the pure cue profiles, excluding memory effects due to memorizing exemplars by their names. To examine the consistency in estimations all profiles were presented twice, again with new names at the second appearance. This allowed us to compare the consistency in estimations for the old profiles encountered in the training phase with the consistency for the new profiles. Larger consistency for known profiles than for new profiles would indicate that memory processes played an important role in the estimations, whereas no differences between old and new profiles would speak in favor of a rule-based approach, described, for instance, by the mapping model. Finally, in Study 2 we aimed for an increased discrimination between the models' predictions.

Method

Participants. In Study 2, 50 participants took part and were randomly distributed to the two conditions, balanced for gender; 25 were women and 25 were men. The mean age was 25 years and the participants were mostly recruited from the universities in Berlin. Participants were paid according to their performance with an average payment of €17 for an individual session lasting on average 1.5 hr.

Design, procedure, and materials. The procedure of Study 2 was similar to that of Study 1, in that participants solved the same estimation task. In contrast to Study 1, the participants only had to learn 19 bugs in the training phase and had to estimate 64 (2×32)

bugs in the test phase. They were told that in the training phase the same 19 bugs would appear 10 times each, whereas in the test phase they would have to evaluate unknown bugs. To ensure that the participants would recognize the bugs when they reappeared, each bug received a male German name. The names were randomly assigned from a list of the most common German names. Otherwise the procedure was the same as in Study 1. Participants were paid according to the accuracy of their estimations. A similar feedback algorithm to that from Study 1 was used, with the correction terms based on the fit of the baseline model (for details see Appendix B).

In Study 2 the training set and the test set were selected in a similar way to Study 1, though with different constraints. The main objective was to improve differentiation between the mapping model, the regression model, and the exemplar model in the linear environment and the mapping model and QuickEst in the J-shaped environment. This was limited, however, by the restriction of unique profiles in the training set. Additionally, in Study 2 the correlation of the cues with the criterion was the same for the linear and the J-shaped environment (but the cue–criterion correlations differed substantially within the environments). Because in Study 1 this correlation differed between the environment conditions, this could explain why the participants differed in their accuracy of estimating the bugs' toxicity in the linear and the J-shaped environment. These changes led to slightly different training sets for the two conditions.

As in Study 1 we examined how well the models predicted the criterion values in the training phase. The exemplar model estimated the criteria perfectly in both environments, due to the unique profiles. All other models did worse with the linear environment than with the J-shaped environment. In the linear environment the regression model was the second-best model, explaining 65% of the variance of the criterion (*RMSD* = 177), whereas the worst

model, QuickEst, explained only 32% of the variance (*RMSD* = 269). In the J-shaped environment, the mapping model reached the second-best accuracy for estimating the criterion values, explaining almost 90% of the variance (*RMSD* = 78), and the regression model was the worst (*RMSD* = 143, r^2 = .60). In sum, the models' accuracies differed substantially for the training phase, which can be explained by two factors. First, we created a task structure that kept the cue–criterion correlations in the linear and the J-shaped environment equal. Second, items were selected such that the differences between the models' predictions for the test phase were increased. Both factors increased the differences of the models' accuracies in the training phase.

Results

Overall, we were able to replicate the results of Study 1. The mapping model was again the best model for predicting participants' estimations when both conditions were considered jointly, and it outperformed all other models in the J-shaped environment. The exemplar model, however, did not substantially profit from the changes in the experimental structure, suggesting that exemplar-based estimation processes do not occur very frequently.

Accuracy and consistency of participants' estimations. The accuracy of the participants was measured in the same way as in Study 1 with the *RMSD* between the participants' estimations and the criterion. The participants mastered the estimation task very easily. The mean *RMSD* dropped in the linear condition from 279 in the first block to 148 in the 10th block. In the J-shaped environment the accuracy increased from an almost equally high error in the first block (*RMSD* = 215) to an *RMSD* of 51 in the 10th block. Just as in Study 1 the data from the three last blocks was merged to analyze the performance. The average *RMSD* in the linear environment was three times as high as in the J-shaped environment [$RMSD_{\text{linear}}$ = 164 vs. $RMSD_{\text{J-shaped}}$ = 58; U = 68, $p < .01$]. Likewise, the

achievement measured by the Pearson correlation between the criterion and the estimations was on average $r = .82$ in the linear environment and $r = .96$ in the J-shaped environment ($U = 87$, $p < .01$). In sum, participants' different accuracies in the two environments reflect the environments' different difficulties.

The cue profiles of the test phase were split into two groups, one consisting of the old profiles known from the training phase and the other containing only new profiles (Table 6). To investigate participants' consistency, the correlations (and the *RMSD*) between the two estimations for the same profile presented twice in the test phase were determined. The participants were equally consistent in the two environments in their estimations for the old profiles, $r_{\text{linear}} = .90$ vs. $r_{\text{J-shaped}} = .89$; $U = 239$; $p = .16$, but the estimations for the new profiles were less consistent in the linear environment than in the J-shaped environment, $r_{\text{linear}} = .67$ vs. $r_{\text{J-shaped}} = .78$, $U = 207$; $p = .04$. The consistency for the new profiles was significantly lower than the consistency for the old profiles, $r_{\text{new}} = .72$ vs. $r_{\text{old}} = .90$; $Z = -5.03$, $p < .01$. The higher consistency in the J-shaped environment indicates that participants relied more on rule-based processes in the J-shaped environment than in the linear environment. However, the drop in consistency from the old profiles to the new profiles suggests memory effects, as the application of rules should not be influenced by the familiarity of the profile.

Response times. In Study 2 we measured the response times for the estimations. Response times dropped during training from a median response of 14.7 s in the first block to 7.5 s in the 10th block. There were no significant difference between the two conditions in the training phase, $Mdn_{\text{linear}} = 8.4$s vs. $Mdn_{\text{J-shaped}} = 7.1$s ($U = 255$, $p = .27$), or the test phase ($U = 238$, $p = .15$). Participants responded faster at the end of the training phase ($Mdn = 7.4$ s) than in the test phase ($Mdn = 9.9$ s; $Z = -4.01$, $p < .01$), but there was no difference in response time between old and new profiles ($Z = -1.3$, $p = .20$).

Model comparison. The fit of the models was quantified in the same way as in Study 1. Again, the data of the last three blocks of the training phase were used to estimate the models' parameters and the fitted models were employed to make predictions for the test phase. Here we focus on the model performance in the generalization test, but the models' fits in the training set can be found in Appendix E. For the generalization test the items of the test phase were split into two groups: one consisting of the old cue profiles encountered in the training phase and the other of only new profiles that had not been encountered before. We first report the results on the old profiles and then come to the decisive comparison in predicting the estimations for the new profiles. In the linear environment, the regression model was the best model for the old profiles, with a significant advantage over the exemplar model ($Z = -2.70$, $p < .01$) and the mapping model ($Z = -3.16$, $p < .01$; see Table 7 for the means). In the J-shaped environment the exemplar and the mapping model were equally good in predicting the estimations for the old profiles in the test phase ($Z = -.71$, $p = .47$) and significantly better than QuickEst or the regression model (mapping model vs. MLR: $Z = -4.32$, $p < .01$).

However, the crucial model test consists of considering how well the models are able to predict participants' estimations for new, independent profiles. As in Study 1, the baseline model was first used as a comparison standard for model performance. For the new profiles, the baseline model reached an average fit of *RMSD* = 213 in the linear environment and of *RMSD* = 136 in the J-shaped environment. Although the exemplar model, the regression model, and the mapping model were better than the baseline model in the linear environment (EBM vs. baseline: $Z = -3.32$, $p < .01$), only the mapping model beat the baseline model in the J-shaped environment (mapping model vs. baseline: $Z = -4.37$, $p < .01$). This indicates that the rather naïve baseline model might not be so bad after all. Especially in the J-shaped

environment, its estimations can be quite accurate, as most of the objects have similarly low criterion values. It also resonates with research on human estimation showing that people tend to rely on the mean if they must predict new objects without further information (Helson, 1964).

When comparing the models against each other the regression model, the mapping model, and the exemplar model were equally good predictors of the participants' estimations of the new objects in the linear environment (see Table 7; mapping vs. MLR: $Z = -.18$, $p = .87$; MLR vs. EBM: $Z = -1.28$, $p = .21$). In the J-shaped environment, the results become much clearer, particularly when we focus on the new objects. The mapping model was the best model; the exemplar model came in second, performing significantly worse than the mapping model ($Z = -3.27$, $p < .01$). Both models performed distinctly better than the regression model or QuickEst. In sum, the two best models (MLR and mapping) demonstrated a quite impressive fit, coming close to the variance in participants' estimations caused by inconsistencies. This error variance provides an upper limit of the fit that can be reached by any deterministic model. Surprisingly, the exemplar model could not predict participants' estimations better than in Study 1, although Study 2 provided better conditions for a memory-based estimation process.

Qualitative analyses. The mapping model proved itself as a valid competitor with the other models. However, to enhance this conclusion drawn from the quantitative model comparison, it is desirable to provide additional qualitative support. The predictions of the mapping model are based on typical criterion values abstracted during the training phase. The mapping model assumes that this typical criterion value is the median criterion value of objects with the same cue sum. Thus the criterion value of some objects in the training set will coincide with the typical criterion value of the mapping model (or be very close to it, if

the median is not defined but the mean of two adjacent objects), while criterion values of others will be clearly different from the typical criterion value. According to the mapping model, objects with criterion values close to the typical criterion value should be estimated more accurately than objects with criterion values differing substantially from the typical value.

In the linear environment, this hypothesis is also compatible with estimations based on the regression model, but in the J-shaped environment the mapping model is the only model that predicts a difference in accuracy between the estimations for objects with typical criterion values and objects with non-typical criterion values. To test this hypothesis, the average errors made on typical objects (objects with the typical criterion value or the two objects with adjacent criterion values) were compared with the errors made on the non-typical objects (all other objects) in the last three blocks of the training set. In both environments the participants made significantly fewer errors estimating the criterion values for objects with typical criterion values than for objects with non-typical criterion values [linear: $RMSD_{\text{typical}}$ = 127, SE = 13; $RMSD_{\text{non-typical}}$ = 179, SE = 17; $t(24) = 22.90$, $p < .01$; J-shaped: $RMSD_{\text{typical}}$ = 38, SE = 6.7; $RMSD_{\text{non-typical}}$ = 54, SE = 7.6; $t(24) = 2.4$, $p = .03$). These results give further support to the mapping model.

Discussion of Study 2

Overall, the results of Study 2 replicated those of Study 1. The mapping model was again best in predicting quantitative estimations, if both environments are considered jointly. In the J-shaped environment, it clearly outperformed the other models. It reached a fit very close to the error variance in the data and was the best model for a distinct majority of participants. In the linear environment, though, it was still not possible to decide unambiguously which model predicted the data of the participants best—the regression

model, the exemplar model, or the mapping model. The differentiation between the models was complicated by the high variance in participants' estimations in the linear environment. In the training phase as well as in the test phase, the estimations showed a high degree of inconsistency. However, the inability of the participants to learn to estimate the criterion values in the linear environment accurately is interesting in itself, as it reflects the poorer ability of the regression model and the mapping model to predict the criterion in the linear environment. Only the exemplar model predicted no differences in learning between the two environments. Because the exemplar model remembers individual cue profiles, its performance is independent of the criterion distribution.

The exemplar model predicted participants' estimations quite well for the old profiles in the test phase, but this was not true for the new profiles. The good fit for the old profiles suggests that participants relied on retrieved exemplars when a cue profile of an object was recognized from the training phase. Unfortunately it does not explain how the estimations for the unknown profiles were made. Here the exemplar model seemed to offer a good description of the estimation process for only a minority of the participants.

Similarly, Juslin et al. (in press) showed in various experiments that the exemplar model described participants' behavior quite well in a "non-linear task," while a regression model was better suited to predict participants' estimations in a "linear task." Similar to our task, the criterion distribution was linear in the linear task and J-shaped in the non-linear task. However, Juslin et al. (in press) conceptualized the difference in the environmental structure not in terms of the distribution of the criterion but by the underlying cue–criterion relationship. The cue–criterion relationship specifies how the criterion is determined as a function of the cue values.

The form of the distribution and the cue–criterion relationship are related in the sense that if representative samples are taken, a linear cue–criterion relationship will result in a roughly linear distribution, and an exponential cue–criterion relationship in a J-shaped distribution. However a linear distribution does not have to stem from a linear function and there are many non-linear functions that would not result in a J-shaped distribution. So far we have not specified the relationship between cues and the criterion in our tasks explicitly but have used a random procedure to generate the criterion distribution. To rule out that this impedes the predictive success of the exemplar model we conducted a third study, in which we chose an approach similar to Juslin et al.'s (in press) to create the objects' criterion values.

Study 3

In Studies 1 and 2 the item sets of the experiments were created by using randomly drawn samples from the simulation study that allowed discrimination between the models. Here the criterion value could only be predicted to some extent by a linear or non-linear function of the cues. Therefore, to further generalize the empirical support for the mapping model, in Study 3 the criterion values were either a linear or a multiplicative function of the cue values (see Juslin et al., in press). Given the results of Studies 1 and 2 we only tested the mapping model against the strongest competing models, which are the standard regression model and the simplified exemplar model with one parameter.

Method

Participants. Forty students from Berlin universities participated in the study, 25 males and 15 females. The mean age was 24 years. The study lasted for approximately 1.5 hr and participants earned on average €17.

Design, procedure, and materials. Study 3 was constructed in the same way as Studies 1 and 2. To generalize the task to further contexts the estimation task was changed to

a medical task. Participants had to estimate the probability that a patient would be cured of a fictitious disease. Participants were told that patients could receive different types of medication and that the information on which drugs a patient took would help them to estimate the criterion, that is, the probability that the patient would be cured within a year, ranging from 1 to 100%. The cues were five different drugs (labeled U, V, W, X, Y), which a patient could either receive or not receive. Participants were told that each drug on its own had a positive effect, but that there could be interaction effects between the drugs. In the linear environment the criterion (C_L) was a linear additive function of the cues (c_i):

$$C_L = 5 + 33c_1 + 22c_2 + 20c_3 + 15\,c_4 + 5c_5$$

In the J-shaped environment the criterion (C_J) was a multiplicative function of the cues:

$$C_J = 1.85 \cdot e^{C_L/25} - 1$$

For a large number of new cases in the generalization test we used a training set of only 16 profiles.

We created 20 different training–test sets that were used for both experimental conditions with 20 participants each. Again we aimed for an experimental item set with large discrimination between the models' predictions. Therefore we first created 1,000 training sets consisting of 16 randomly selected cue profiles and by using the two functions we determined the criterion values. The respective generalization sets consisted of the 16 profiles that did not appear in the corresponding training set. Next we excluded all sets in which cues correlated negatively with the criteria. Then we rank ordered the training sets according to how well they discriminated between each possible pair of models in the generalization set and chose the 20 environments that allowed maximum discrimination between all models. The experimental procedure was the same as that used in Studies 1 and 2. During a training phase consisting of 160 trials, participants learned to estimate the criterion value connected with

each profile in the training set. After each trial participants received feedback on the correct criterion values and their performance. The order of appearance was randomized as well as the assignment of the cues to the five different drugs and the order in which the drugs appeared on the screen. In the test phase each participant estimated all possible profiles two times without feedback. Participants were paid according to a feedback algorithm that was determined in the same way as in Studies 1 and 2 (for details see Appendix B).

Results

Overall, Study 3 replicated the results of Studies 1 and 2. The mapping model was clearly the best model in the J-shaped environment. However, in the linear environment the regression model outperformed the other models. Before we come to the model comparisons we report the participants' accuracy.

Accuracy of the participants.

The participants learned to estimate the criterion quite well in both conditions. In the linear environment the *RMSD* dropped from 18 in the first block to 7 in the 10th block and in the J-shaped environment from 29 to 7, with a rather stable accuracy in the last three blocks of the training phase. Participants' accuracy at the end of the training phase did not differ significantly between the two environments ($M_{\text{Linear}} = 7$ *RMSD* vs. $M_{\text{J-shaped}} = 8$ *RMSD*; $U = 176$, $p = .53$).

Did the participants capture the underlying function generating the criterion values? This can be seen in how well participants could predict the criterion values of the new cue profiles in the generalization set. In both environments participants were worse at estimating criterion values of patients with new drug combinations than with previously encountered combinations ($M_{\text{old}} = 7$ *RMSD* vs. $M_{\text{new}} = 14$ *RMSD*, $Z = -5.3$, $p < .01$). However, they were significantly better in the linear environment than in the J-shaped environment ($M_{\text{J-shaped}} = 16$

vs. $M_{\text{Linear}} = 12$; $U = 123$, $p = .04$). This suggests that the participants in the linear environment captured the function generating the criterion values to some extent.

Model comparison. As in the preceding studies, the models were fitted on the last three blocks of the training phase for each. For the crucial model comparison test we focused on the generalization test of the test phase. In particular we compared the accuracies of the models in predicting participants' estimates of the criterion values for the new cue profiles, that is, combinations of drugs they had not seen during the training phase. Here the results were clear-cut. In the linear environment the regression model predicted participants' estimations significantly better than all other models, with the mapping model coming in second ($M_{\text{MLR}} = 9$ *RMSD* vs. $M_{\text{mapping model}} = 14$ *RMSD*, $Z = -3.1$, $p < .01$). In the J-shaped environment the mapping model clearly outperformed all other models ($M_{\text{mapping model}} = 10$ *RMSD* vs. $M_{\text{MLR}} = 17$ *RMSD*, $Z = -3.3$, $p < .01$). The exemplar model and the regression model performed equally poorly. Figure 2 illustrates the accuracies of the different models in predicting participants' estimations in the generalization test.

Discussion of Study 3

We conducted Study 3 to test if our results from Studies 1 and 2 would also hold if the criterion distributions were generated by a linear and a non-linear function of the cue values. In the J-shaped environment this was clearly the case. In the linear environment, however, linear regression outperformed the mapping model. As the linear criterion was generated by a linear additive function, the regression model could now be equivalent to the function generating the criterion values and could estimate the criterion faultlessly. Thus if participants were able to detect the underlying structure in the data, then the regression model would capture their estimations. We will discuss this issue further in the General Discussion.

In the J-shaped environment, we did not find a shift to an exemplar-based estimation process as advocated by Juslin et al. (in press); instead, the mapping model still described participants' behavior best. This corroborates that the mapping model is the best model for J-shaped distributions regardless of whether the underlying function has been specified.

Study 4

Study 4 represents a reanalysis of Experiment 1 of Juslin et al. (in press). In this study the authors found empirical support for a rule-based estimation process in an environment with a linear distribution of the criterion, whereas support for an exemplar-based estimation process was reported for an environment with a J-shaped criterion distribution. To test whether the mapping model, which Juslin et al. did not examine, offers an alternative account of the estimation processes, we reanalyzed the experimental data.

Juslin et al.'s experiment differed in important aspects from our studies. First, in the training phase of the experiment the participants were confronted with only 11 different profiles, a small number, that were described by only four cues, as opposed to 20, 19, and 16 different profiles with five cues each in Studies 1, 2, and 3, respectively. Second, although participants had to process less information in the training phase compared to our studies, much more training was provided by repeating each profile 20 times, as opposed to 10 repetitions in our studies. This procedure should have made it easier to memorize each profile, thus fostering an exemplar-based estimation process. Moreover, and maybe most importantly, in the experiment by Juslin et al. the participants had to learn the direction of the cues during the training phase, while in our studies the direction of the cues was told to the participants. Additionally, the cue–criterion correlations of some cues were rather small and fluctuated during training, increasing the difficulty of learning the correct direction of the cues.[6] We think this last difference is disadvantageous for a rule-based estimation process, as

described by the mapping model, for which the cue–criterion correlations are essential. In sum, we think the experimental procedure is beneficial for an exemplar-based estimation process and it would be surprising if the mapping model still predicted people's behavior well.

Method

Design and procedure. The experiment had two conditions in which participants had to estimate the toxicity of bugs based on four binary cues, similar to Studies 1 and 2. In the first condition, the linear condition, the criterion was a linear function of the cues; in the second, the multiplicative condition, the criterion was a multiplicative function of the cues. Similar to our tasks the criterion values followed either a linear or a J-shaped distribution. In an initial training phase with 220 trials participants learned to estimate the criterion values on a subset of 11 of the 16 possible bugs. In a subsequent test phase they then estimated the toxicity of all 16 bugs, that is, including the 5 bugs that they had not encountered before.

Model fitting. Following the same procedure used by Juslin et al. (in press) we fitted the models on the second half of the training data. Juslin et al. used the standard exemplar model with a free parameter for each cue, so we included this version together with the simplified exemplar model that we have reported so far. Thus, we will report results for two exemplar models, one complex exemplar model with four free parameters and one simple exemplar model with one free parameter. As in our preceding studies we analyzed the data on the individual level. We estimated the exemplar models' parameters on the second half of the training set starting with a memory base consisting of the correct cue and criterion values of the first half of the training set.[7] Then we successively added the exemplars of the second half of the training set to the memory base in the order in which they were encountered. This way the memory base always represented all objects the respective participant had seen so far (we

think this method is most appropriate, because due to random error the same cue profiles had varying criterion values). The regression model was fitted directly to the participants' estimations from the second half of the training set. Consistent with our previous studies but in contrast to Juslin et al., we used an unconstrained linear regression.[8] For the mapping model we determined the directions of the cue–criterion relationship on the basis of the correlation of the cue with the criterion in the second half of the training set and then calculated the typical criterion values for each cue sum category based on the criterion values. With the estimated parameters from the training phase, each model predicted estimations for the test phase.

Results

Overall, we replicated the results of Juslin et al. (in press), but our results were not quite as clear-cut. The regression model performed best in the linear condition and the exemplar model with one parameter was the best model in the multiplicative condition. However, the simplified exemplar model was not significantly better than the regression model and the mapping model performed as well as the standard version of the exemplar model.

Model comparison. Surprisingly, and in contrast to the results of our Studies 1–3, all models performed worse in the training phase than in the test phase. In the test phase, the regression model performed best in the linear condition. It was significantly better than the mapping model and the simplified exemplar model. However, the comparison between the regression model and the standard exemplar model with four parameters only approached significance, $M_{\mathrm{MLR}} = 1.4$ *RMSD* vs. $M_{\mathrm{EBM}} = 1.5$ *RMSD*, $Z = -1.78$, $p = .08$.

In the multiplicative condition it was difficult to identify one best model. The standard exemplar model used by Juslin et al. (in press) was statistically indistinguishable from the

mapping model, the regression model, and the simplified exemplar model. However, the simplified exemplar model with one parameter performed slightly better than the regression model (M_{MLR} = 1.8 *RMSD* vs. M_{EBM} = 1.7 *RMSD*, $Z = -1.65$, $p = .10$) and was significantly better than the mapping model ($M_{mapping}$ = 2.0 *RMSD*, $Z = -3.1$, $p < .01$).

Discussion of Study 4

In contrast to Studies 1–3, the mapping model performed as well as or worse than the linear regression or the simplified exemplar model. This result highlights the dependence of the models' predictive accuracy on the structure of the task and indicates boundary conditions for the mapping model.

In the linear condition participants were able to pick up the linear additive structure of the task. Thus, in line with the reasoning of Juslin et al. (in press) and the results of Study 3, the regression model was the best model in the linear condition. In the multiplicative condition, however, the simplified exemplar model described participants' behavior better than the mapping model. We assume that this difference is due to the experimental procedure employed by Juslin et al., which was different from that employed in our studies. Due to a smaller number of cue profiles and more extensive training, memorization of exemplars was presumably enhanced, favoring an exemplar-based estimation process. In contrast, the mapping model was constrained by the small number of cues and the selection of the training examples. Due to the composition of the training set the mapping model could only establish three categories, limiting the number of possible estimates to a rather small number.

However, the presumably most important difference in the tasks lies in the correlation of the cues with the criteria. In the experiment by Juslin et al. (in press), the direction of the cues had to be detected by the participants. The mapping model assumes that knowledge about the correct cue directions can be learned from the environment, but it does not specify

the learning process. Thus, we assumed that participants picked up the cues' directions from the training set. However, as some cue–criterion correlations were rather small, it could easily be that some participants got the direction of the cues wrong or ignored cues that did not seem predictive for the task. In such a situation—where the direction of the cue–criterion correlation is unclear, participants have extensive experience with the exemplars, and the criterion is a non-linear function of the cues—a shift to an exemplar-based process seems plausible. However, if all cues reliably predict the criterion and their direction is known to the participants, the mapping model seems to be the better model.

General Discussion

To describe the cognitive process underlying quantitative estimations we proposed a new cognitive theory that we called the mapping model. In four studies we tested this model against three alternative estimation models under a variety of environment conditions. We examined how well the models predicted estimations in a linear environment with a linear additive cue–criterion relationship, as opposed to a J-shaped environment with a non-linear cue–criterion relationship.

The Success of the Mapping Model

The mapping model is built on an existing, successful framework for quantitative estimations—the so-called metrics and mappings framework (Brown & Siegler, 1993, 1996; Brown, 2002). Implementing a computational model of this framework enabled us to test the mapping model against other cognitive computational models of estimations. Naturally the way we specified the mapping model is only one possibility and there might be other and better ways to do so. Nevertheless, we think that our model captures the general idea of the metrics and mapping framework, and when considering the empirical evidence provided by Studies 1–3, the model appears successful in predicting people's estimations. In three out of

four studies, the mapping model was clearly superior to the other models in the J-shaped environment. Even in the linear environment, where a clear advantage of linear regression might have been expected, it performed equally well and was only outperformed when the criterion was perfectly predictable by a linear regression.

Rule-based Estimation

In the J-shaped environments the regression model was clearly not the appropriate model to predict participants' estimations. In the linear environments, the results were less clear. The regression model predicted participants' estimations as well as the mapping model in the first two studies but outperformed the mapping model in Studies 3 and 4. This resonates with innumerable articles that have shown that the regression model can successfully capture linear judgments (see Hammond & Stewart, 2001; Brehmer & Brehmer, 1988).

The varying results might be explained by an adaptive response to the environment (for a theoretical account see Rieskamp, Busemeyer, & Laine, 2003; Rieskamp & Otto, 2006). Because in Studies 3 and 4 the criterion values were generated by a linear additive function of the cues, the regression model was the optimal model for predicting the criterion. Thus, in their attempts to be behave adaptively, the participants might have learned to follow a linear additive estimation strategy, as captured by the regression model. This adaptive response to the environment was also enhanced by the ease with which optimal cue weights of a linear additive estimation strategy could be abstracted during training. In Study 3 optimal cue weights could be reliably estimated from any pair of objects differing on only one cue. That is, when the cue changed from a negative to a positive cue value from one object to another, the criterion value always increased by a constant amount. In Studies 1 and 2, in contrast, an estimation process in line with the mapping model was equally successful. The

regression model could only approximately predict the criterion and it was more difficult to judge a cue's contribution correctly. This might have favored an approach of giving equal weights to all cues, as assumed by the mapping model. It could also help explain why the regression model and the mapping model could not be distinguished in Studies 1 and 2. In a linear environment the mapping model can be equivalent to a unit weight regression model, so that the systematic variance captured by the mapping model and the regression model potentially overlap. In addition, research on linear regression models has often shown a flat maximum effect, where equal weights lead to the same accuracy in prediction as optimal weights (Dawes, 1979; Einhorn & Hogarth, 1975).

In sum, in a task where the criterion is a linear additive function of the cues, people appear to be able to recognize the structure underlying the data and to abstract the appropriate weights for a linear additive estimation process. Consequently participants' estimations in such a situation are best described by the regression model. However, when abstracting the optimal cue weights is complicated, because the criterion is not a linear additive function of the cues, a shift to a unit weight model such as the mapping model seems to take place.

Exemplar-based Estimations

Research by Juslin et al. (2003b, in press) suggests that in the case of a linear cue–criterion relationship, rule-based processes offer a better description of human estimation than exemplar-based models. Consistently the regression model or the mapping model was best in predicting estimations when the criterion values were linearly distributed. Consistent with Juslin et al.'s (in press; Karlsson, Juslin, & Olsson, 2004) prediction that exemplar-based processes should occur for non-linear cue–criterion relationships, we found that the exemplar model outperformed the regression model in predicting participants' estimations in J-shaped environments. However, in three of the four studies the mapping model as opposed to the

exemplar model was best in predicting participants' estimations and only in Study 4 was the exemplar model best. Thus other factors besides the criterion distribution or the functional cue–criterion relationship seem to drive the models' predictive success.

The number of exemplars as well as the number of cues on which the exemplars differ and the amount of experience needed to memorize exemplars appear important: The exemplar model requires that all or at least a majority of the objects encountered during training be stored. Therefore the more information there is that has to be stored and the less often each object is encountered, the more difficult memorizing complete exemplars should become. If memorization of exemplars is difficult, a shift to a less demanding estimation process, captured by the mapping model, should be expected. Consistently we found that the mapping model performed better in Studies 1–3.

Further, the direction and the magnitude of the cue–criterion correlations and how reliably they can be abstracted when gaining experience with an estimation situation could influence the models' predictive success. For the exemplar model the direction and the magnitude of the cue–criterion correlation is not decisive. As long as objects can be sufficiently differentiated by their cue profiles, the exemplar model will always reach perfect performance with a given set of known objects. In contrast, the mapping model relies on knowing the correct direction of the cues. However, the mapping model does not specify how knowledge about the cues' direction is acquired, but we made the simplifying assumption that participants learn the correct direction during training. This appears justified when the cues correlate substantially with the criterion. In Study 4, however, some of the cues did not predict the criterion very well, with cue–criterion correlations fluctuating around zero, making it difficult to detect the cues' directions. This indicates that in a situation where it is difficult to abstract the direction of the relationship of the cues with the criterion and where

the quality of the cues is dubitable, the exemplar model might be more suitable for predicting estimation processes than the mapping model.

In sum, the characteristics of the estimation situation of Study 4 were beneficial for an exemplar-based estimation process and detrimental for a rule-based process. We identified two task factors that influence the predictive success of the mapping model and the exemplar model in predicting estimations. We expect that the mapping model will be able to predict estimations in situations where many predictive cues are available, prior knowledge about the cues exists, and training is short. The exemplar model will be better in situations where the quality and the direction of the cues is unclear and extensive training on objects differing only on a few predictive cues is available. These expectations require further empirical tests.

Simple Heuristics for Estimation

In Study 1 a considerable number of participants were best described by QuickEst in the J-shaped environment. This raises the question of under which the conditions QuickEst might capture the process of human estimation. QuickEst does not integrate information, whereas the mapping model uses all information available. For probabilistic inference tasks it has been found that models integrating information are often good predictors of people's inferences when all information is visible and easily accessible (Bröder, 2000; Newell & Shanks, 2003). In contrast, when information search is costly, shifts to simple heuristics that do not integrate information have been reported (e.g., Payne, Bettman, & Johnson, 1993; Rieskamp & Hoffrage, 1999; Rieskamp & Otto, 2006). This suggests that QuickEst was in a disadvantageous position in our experiments, in which all information was presented simultaneously on the screen. Another recent study has also not found any empirical support for QuickEst (Hausmann, Läge, Pohl, & Bröder, in press).

Complexity of the Models

The models we considered differed in their complexity, that is, their flexibility in predicting different behaviors. Though complex models are better in fitting data, they face the problem of over-fitting—instead of describing the systematic structure of the cognitive process underlying estimation they might fit unsystematic error. To reduce the problems of model complexity in model selection we relied on a generalization test and included simplified versions of the models. To our surprise the complex standard regression model, with a free parameter for each cue, performed better than the simplified versions of the regression model. Thus, only by using its full complexity was the regression model able to predict people's estimations.

However the standard exemplar model (Medin & Schaffer, 1978, adapted by Juslin et al., 2003b), with one free parameter for each cue, apparently over-fitted the data. Overall, the simplified exemplar model, assuming that all cues are regarded as equally important, provided a better account of people's estimations. Thus the psychological interpretations of the attention parameters of the original exemplar model representing the subjective importance of each cue should be treated very cautiously. Further the linear condition of Study 4 indicates that there might be inference situations in which it becomes necessary to assume different attention weights for the cues. This leaves open the problem of predicting a priori which of the two exemplar models will predict behavior better.

The mapping model was the simplest model we considered as it entailed no free parameters and we only tested one version of it. Without flexibility, the model is unable to capture any specific ways people respond to a particular estimation situation. However, this disadvantage can turn out to be an advantage: the lack of flexibility reduces the danger of over-fitting, thereby making predictions more robust. This is particularly important because

the environments we encounter in everyday life are typically noisy. For instance, environments can rarely be expressed by a linear additive function of cues, which could favor the unit weight approach taken by the mapping model. In a similar vein, the mapping model reduces the information load by ignoring the pattern of the cue values. In environments where it is unclear which cues can help to predict the quantity of interest, this might not be a utile assumption. However, if a set of predictive cues has been identified, the assumption appears psychologically plausible, making the mapping model a good model of human estimations.

Limitations of the Mapping Model

What are the boundary conditions of the mapping model's success in predicting quantitative estimations? For one, we showed that the mapping model can be outperformed by linear regression when the criterion is a linear additive function of the cues. This touches upon a limitation of the mapping model: It relies exclusively on the objects it has encountered so far, so that—in contrast to the regression model—it is unable to extrapolate over the range of encountered criterion values. Research on function learning has shown that with sufficient practice, people are quite adept at learning a variety of one-dimensional functions (Kalish, Lewandowsky, & Kruschke, 2004; DeLosh, Busemeyer, & McDaniel, 1997; for a review see Busemeyer, Byun, DeLosh, & McDaniel, 1997). However, if multiple cue dimensions have to be integrated into a single response, the ability to extrapolate seems to be restricted to linear functions. Juslin et al. (in press; see also Karlsson et al., 2004; Olsson et al., 2006) showed in several studies that participants did not extrapolate if the cues were non-linearly connected to the criterion. Thus, we believe that the mapping model's inability to extrapolate could to some extent mirror human behavior.

The comparison with the exemplar model in Studies 3 and 4 highlighted other boundary conditions for the mapping model. The model assumes that the direction of the cue–

criterion relationship can be learned from the environment. When this is complicated the cues' direction assumed by the mapping model might not correspond with the subjective directions perceived by a decision maker, so that the predictions of the mapping model become inaccurate. Likewise, the mapping model does not specify which cues are used for the estimation process but includes all cues. In a condition where all cues are good predictors of the criterion this is a reasonable strategy, but in situations in which a few good predictors have to be picked out of a bunch of irrelevant cues, it will not work well. Further, the mapping model relies on a representative sample of criterion values for each category. In the case where a cue sum category is only represented by an outlying criterion value of one single object, the estimation of the mapping model can be distorted. Finally, another limitation of the mapping model is its application to estimation problems with only binary cues. How can the mapping model be extended to continuous cues? One way would be to dichotomize continuous cues (e.g., by a median split). However this rather crude approach might result in an overly strong loss of information. A second possibility would be to reduce a large number of cue sum categories to a few manageable categories, for instance, by applying range–frequency theory (Parducci, 1965).

Final Conclusion

Past research on multiple cue judgments has focused on linear regression as a tool to analyze human judgments (Brehmer, 1994; Hammond, 1996). Although linear additive models can predict the outcome of estimation processes rather well, they have been criticized for not capturing the underlying cognitive process (e.g., Gigerenzer & Todd, 1999; Hoffman, 1960; see also Doherty & Brehmer, 1997). In response to this criticism, alternative estimation models have recently been proposed and tested, including exemplar models adapted to estimation problems (Juslin et al. 2003b; Medin & Schaffer, 1978), and simple heuristics such

as QuickEst (Hertwig et al., 1999). Following up on the criticism, we proposed the mapping model as a simple, new cognitive theory and showed that it can successfully predict human estimation.

References

Albers, W. (2001). Prominence theory as a tool to model boundedly rational decisions. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (pp. 297–317). Cambridge, MA: MIT Press.

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

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87,* 137–154.

Brehmer, A., & Brehmer, B. (1988).What have we learnt about human judgment from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75–114). Amsterdam: Elsevier/North Holland.

Brehmer B., & Joyce, C. R. B. (Eds.). (1988). *Human judgment: The SJT view.* Amsterdam: Elsevier/North Holland.

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.

Brown, N. (2002). Real world estimation: Estimation modes and seeding effects. In B. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 321–359). San Diego, CA: Academic Press.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*, 511–534.

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin and Review, 3*, 385–388.

Brunswik. E. (1952). *Conceptual framework of psychology*. Chicago: University of Chicago Press.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Busemeyer, J. R., & Wang, Y-M. (2000). Model comparisons and model selection based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' predictions of children's early reading achievement: An application of social judgment theory. *American Educational Research Journal, 23*, 41–65.

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

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 968–986.

Doherty, M., & Brehmer, B. (1997).The paramorphic representation of clinical judgment: A thirty-year retrospective. In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections and controversies* (pp. 537–551). Cambridge: Cambridge University Press.

Doherty, M. E., & Kurz, E. (1996). Social judgement theory. *Thinking and Reasoning, 2*, 109–140.

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

Ebbesen, E. B., & Konecni, V. J. (1975). Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology*, *32*, 805–821.

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

Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review, 86,* 465–485.

Fishbein, M., & Ajzen, I. (1980). *Understanding attitudes and predicting social behavior*. New York: Prentice Hall.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, *114*, 739–767.

Gigerenzer, G., & Kurz, E. (2001). Vicarious functioning reconsidered: A fast and frugal lens model. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik:*

Beginnings, explications, applications (pp. 342–347). New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

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

Gonzalez-Vallejo, C., & Bonham, A. (in press). Aligning confidence with accuracy: Revisiting the role of feedback. *Acta Psychologica*.

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

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Harries, P. A., & Harries, C. (2001). Studying clinical reasoning. Part 2: Applying social judgment theory. *British Journal of Occupational Therapy, 64*, 285–292.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining inference and prediction*. New York: Springer.

Hausmann, D., Läge, D., Pohl, R., & Bröder, A. (in press). Testing the QuickEst: No evidence for the Quick-Estimation heuristic. *European Journal of Cognitive Psychology*.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P.M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hertwig, R., Hoffrage, U., & Sparr, R. (2007). *The QuickEst heuristic: How it benefits from an imbalanced world.* Manuscript in preparation.

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

Hogarth, R. M., Gibbs, B. R., McKenzie, C. R. M., & Marquis, M. A. (1991). Learning from feedback: Exactingness and incentives. *Journal of Experimental Psychology: Learning Memory and Cognition, 17*, 734–752.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003a). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*, 924–941.

Juslin, P., Karlsson, L., & Olsson, H. (in press). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*.

Juslin, P., Olsson, H., & Olsson, A-C. (2003b). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133–156.

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563–607.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111*, 1072–1099.

Karlsson, L., Juslin, P., & Olsson, H. (2004). Representational shifts in a multiple-cue judgment task with continuous cues. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 648–653). Mahwah, NJ: Cognitive Science Society.

Levy, M., & Solomon, S. (1997). New evidence for the power-law distribution of wealth. *Physica, 242*, 90–94.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Myung, J. I., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 422–437), London: SAGE Publications.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 53–65.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes* (Vol. 1, pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*, 375–402.

Olsson, A. C., Enqvist, T., & Juslin, P. (2006). Go with the flow! How to master a nonlinear multiple-cue judgment tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1371–1384.

Olsson, H., Wennerholm, P., & Lyxzén, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning , Memory, and Cognition, 30*, 936–941.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407–418.

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

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 1371–1384.

Rieskamp, J., Busemeyer, J. R., & Laine, T. H. (2003). How do people learn to allocate resources? Comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *29*, 1066–1081.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics and how can we tell? In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 141–167). New York: Oxford University Press.

Rieskamp, J., & Otto, E. P. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

Schroeder, M. (1991). *Fractals, chaos, power laws: Minutes from an infinite paradise.* New York: Freeman.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3–21.

Wigton, R. S. (1996). Social judgement theory and medical judgement. *Thinking and Reasoning, 2*, 175–190.

Wryobeck, J. M., & Rosenberg, H. (2005). The association of client characteristics and acceptance of harm reduction: A policy capturing study of psychologists. *Addiction Research and Therapy, 13*, 461–476.

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behaviour and Human Performance, 18*, 269–294.

Appendix A

Simulation Procedure

The simulation study examined in a 9 × 4 (in the J-shaped environment) and 9 × 6 (in the linear environment) factor design the impact of the percentage of negative cue values and the magnitude of the cue–criterion correlation. The conditions for the simulation were created in several steps. First, nine sets of five dichotomous cues with differing percentages of negative cue values were created. All cues of a set shared the same percentage of negative cue values, varying in steps of .10, between .10 and .90 per cue. The cue values were randomly assigned to the 50 objects representing an environment. Second, for each level of percentage of negative cue values we created further sets to manipulate the cue–criterion Pearson correlation. For each set with the same percentage of negative cue values we created different sets with different cue–criterion correlations. The cue–criterion correlations were varied in steps of .10 between .0 and .30 in the J-shaped environment (providing four different levels) and between .0 and .50 in the linear environment (providing six different levels). Again, all cues of a set shared the same correlation. Because the maximal possible correlation decreases with increasing percentages of positive cue values in the J-shaped environment, the number of factor levels for the correlations was lower in the J-shaped environment. The cue–criterion correlations were modified by randomly selecting two objects with different cue values and exchanging their cue values if this changed the cue–criterion correlation in the desired direction (this step was repeated until the desired correlations were obtained). This resulted in a 9 (percentage of negative cue values) × 4 (magnitude of correlation) design in the J-shaped environment and a 9 (percentage of negative cue values) × 6 (magnitude of correlation) design in the linear environment. In every condition each model was fit to half of the data and then tested on the other half.

Appendix B

Feedback Algorithms

During the training phase participants got feedback about the accuracy of their estimations and the number of points they were earning. The points participants received for their estimations were determined by the following algorithms. Any unusual deviation exceeding 500mg/l, as might be caused by a typing mistake, was treated as a deviation of 500mg/l. For each environment a different correction term (e.g., 1,100 for the linear environment in Study 1) was used to adjust for the task difficulty. The correction term was chosen dependent on the baseline model and determined the magnitude of the deviation for which a participant would receive zero points. The magnitude of the deviation that would result in zero points is given by the root of the correction term multiplied by 100. Thus in the linear environment a participant deviating less than $332 = (1{,}100 \times 100)^{1/2}$ mg/l would earn points whereas for a deviation exceeding 332 mg/l, points would be subtracted.

The equations for the feedback algorithms are defined as

$$y = -x^2/c + 100, \text{ for } x \leq 500 \text{ and}$$

$$y = -500^2/c + 100, \text{ for } x > 500,$$

where x is the absolute difference between a participant's estimation and the actual criterion value for a given trial, y denotes the number of points that were added or subtracted from the participant's account, and c is the correction term. The correction terms for Study 1 were c = 1,100 for the linear environment and c = 550 for the J-shaped environment. The correction terms for Study 2 were c = 888.58 for the linear environment and c = 536.26 for the J-shaped environment. The correction terms for Study 3 were c = 556 for the linear environment and c = 512 for the J-shaped environment.

Appendix C

Structure of the Test Sets in Study 1

In Study 1 the test sets in the two environments differed slightly. Each test set consisted of old objects that had appeared in the training phase and new objects that the participants had not encountered before (see Tables C1 and C2).

Table C1

Test Set in the J-shaped Environment in Study 1

Number	Profile	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Exemplar	Regression	QuickEst	Mapping
1	Old	0	0	0	0	0	23	37	30	23
2	Old	0	0	0	1	0	33	25	30	40
3	Old	0	0	1	0	1	34	32	30	34
4	Old	0	1	0	0	0	75	61	50	40
5	Old	0	1	0	1	0	41	49	50	34
6	Old	0	1	0	1	1	130	131	70	71
7	Old	0	1	1	0	1	52	56	50	71
8	New	0	1	1	1	1	284	44	70	286
9	New	1	0	0	0	0	23	559	30	40
10	New	1	0	0	0	1	29	641	30	34
11	New	1	0	0	1	0	33	548	30	34
12	New	1	0	0	1	1	232	629	30	71
13	New	1	0	1	0	1	34	554	30	71
14	New	1	0	1	1	1	566	543	30	286
15	New	1	1	0	0	0	75	584	50	34
16	New	1	1	0	0	1	242	665	50	71
17	New	1	1	0	1	0	42	572	50	71

Number	Profile	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Exemplar	Regression	QuickEst	Mapping
18	New	1	1	0	1	1	317	653	500	286
19	New	1	1	1	0	1	438	579	50	286
20	New	1	1	1	1	0	566	485	50	286
21	Old	1	1	1	1	1	567	567	500	500

Note. The profiles are ordered lexicographically according to the cues' correlation with the criterion in the training set. Profiles 1–7 and 21 also appeared in the training set. The parameters for the models were set as follows: Exemplar model with one free parameter: s = .0006; regression model: intercept = 36.92, c_1 = 522.39, c_2 = 24.16, c_3 = −86.23, c_4 = −11.83, c_5 = 81.25; for QuickEst all cues were included.

Table C2

Test Set in the Linear Environment in Study 1

Number	Profile	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	Exemplar	Regression	QuickEst	Mapping
1	Old	0	0	0	0	0	50	151	200	50
2	Old	0	0	0	1	0	220	150	200	300
3	Old	0	0	1	0	1	240	244	200	240
4	Old	0	1	0	0	0	480	379	300	300
5	Old	0	1	0	1	0	307	377	300	240
6	Old	0	1	0	1	1	665	665	700	640
7	Old	0	1	1	0	1	480	472	700	640
8	New	0	0	1	1	1	240	243	200	640
9	New	0	1	1	1	1	738	470	700	780
10	New	1	0	0	0	1	145	889	200	240
11	New	1	0	0	1	0	220	600	200	240
12	New	1	0	0	1	1	608	888	200	640
13	New	1	0	1	0	1	240	69	200	640
14	New	1	0	1	1	1	920	693	200	780
15	New	1	1	0	0	0	480	829	300	240
16	New	1	1	0	0	1	686	1117	700	640
17	New	1	1	0	1	0	307	827	300	640
18	New	1	1	0	1	1	774	1115	700	780
19	New	1	1	1	0	1	810	922	700	780
20	New	1	1	1	1	0	920	632	300	780
21	Old	1	1	1	1	1	920	920	700	920

Note. The profiles are ordered lexicographically according to the cues' correlation with the criterion in the training set. Profiles 1–7 and 21 also appeared in the training set. The parameters for the models were set as follows: Exemplar model with one free parameter: s =

.0001; regression model: intercept = 151.32, c_1 = 450.09 c_2 = 227.37, c_3 = −195.09, c_4 = −1.67, c_5 = 287.98; for QuickEst all cues were included.

Appendix D

Comparison of the Standard Exemplar Model, the Simplified Exemplar Model, and the Regression Model

In Study 1 we included simplified variants of the regression model and the exemplar model in addition to the standard versions. For the exemplar model we included a version with five parameters fit to each participant individually (standard exemplar), a simplified exemplar model with only one free parameter (simplified exemplar), and an exemplar model with its five parameter values optimized by using the objective criterion value instead of participants' estimations (a priori exemplar). For the regression model we included the standard model with six free parameters fit to each participant individually (standard regression), a stepwise regression model that only included the cues that received significant weights (stepwise regression), and a regression model with the parameter values optimized by using the objective criterion value instead of participants' estimations (a priori regression).

The parameters of the simplified variants of the exemplar model and the regression model were estimated in the same way as for the standard versions. For the a priori exemplar model and the a priori regression model the parameters were optimized by using the objective criterion values of the training set. The simplified exemplar model and the standard exemplar model were fitted on the last three blocks of the training phase with the correct cue and criterion values of the training set as the memory base. The best parameters for each participant were searched for by using the quasi-Newton optimization method as implemented in MATLAB. To avoid local minima, parameters were first derived by a grid search with the results serving as the starting values for the subsequent fitting procedure. The parameters for the standard regression model and the stepwise regression model were obtained by respectively determining a multiple linear regression and a stepwise regression

on the last three blocks of the training set. The stepwise regression model reduced the number of employed cues substantially; on average only 3.7 (SD = 1.06) cues were used in the linear environment and only 1.5 (SD = .77) in the J-shaped environment.

Naturally, of the different versions of the exemplar and regression models, when fitted to the data of the training phase the most complex ones did significantly better than the simplified versions. In the crucial generalization test of the test phase, however, the simplified exemplar model was clearly superior to the standard version of the exemplar model and the a priori exemplar model (all Zs < −2.48, ps < .01). The standard version of the regression model in all cases did significantly better than the two simplified versions except in the J-shaped environment, where the a priori regression model was equally as good as the standard regression model (Z = −.59, p = .57). Here we report the $RMSD$s of all versions for the test phase (see Table D1).

In Study 2 we again tested all versions of the exemplar model and the regression model in the model comparison. But similar to in Study 1, the stepwise regression and the regression with the parameters set a priori performed worse than the full model. The simplified exemplar model also performed again significantly better than the standard exemplar model and the a priori exemplar model.

Table D1

Average Predictive Accuracy of the Models in the Test Set of Study 1

	Standard exemplar	Simplified exemplar	A priori exemplar	Standard regression	Stepwise regression	A priori regression
			Linear environment			
RMSD	219	161	206	166	182	282
SD	60	40	37	56	58	45
			J-shaped environment			
RMSD	242	166	179	342	359	352
SD	89	70	73	124	123	72

Note. The J-shaped environment condition had 30 participants and the parameters determined a priori for the exemplar model were s_1= .0055, s_2 = .0008, s_3 = .0088 s_4 = .0005, and s_5 = .0006; the parameters for the regression model were intercept = 36.92, c_1= 522.39, c_2 = 24.16, c_3 = −86.23, c_4 = −11.83, and c_5 = 81.25. The linear environment condition had 29 participants and the parameters determined a priori for the exemplar model were s_1= .0274, s_2 = .0002, s_3 = .0049, s_4 = .0001, and s_5 = .0001; the parameters for the regression model were intercept = 151.32, c_1= 450.09, c_2 = 227.37, c_3 = −195.09, c_4 = −1.67, and c_5 = 287.98.

Appendix E

Model accuracies for the training phase of Study 2.

In Study 2 all models performed better than the baseline model in predicting participants' estimations for the training phase. Because the training phase consisted of unique profiles, we expected the exemplar models to reach a fit close to the participants' accuracy. As anticipated, the exemplar model performed very well, explaining over 74% of the variance in the linear environment and 90% in the J-shaped environment. The models' accuracies are reported in Table E1.

Table E1

Model Accuracies in the Training Set of Study 2

	Environment							
	Linear				J-shaped			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
RMSD	192	153	253	138	83	150	144	56
SD	30	28	18	54	18	11	19	35
r^2	.58	.71	.31	.74	.88	.58	.75	.92
SD	0.12	0.09	0.06	0.14	0.05	0.06	0.11	0.08

Authors' Note

Bettina von Helversen and Jörg Rieskamp, Max Planck Institute for Human Development, Berlin, Germany. We would like to thank Peter Juslin and Linnea Karlsson for providing us with the experiment data of their previous work for a re-analysis. We gratefully acknowledge helpful comments on previous versions of this article by Peter Frensch, Peter Juslin, and Konstantinos Katsikopoulos. We would like to thank Anita Todd for editing a draft of this manuscript. Correspondence concerning this article should be addressed to Bettina von Helversen.

Bettina von Helversen

Max Planck Institute for Human Development

Lentzeallee 94, 14195 Berlin, Germany

Phone: (+49 30) 82406 699

Fax: (+49 03) 82406 394

Email: vhelvers@mpib-berlin.mpg.de

Footnotes

1. We chose the median as opposed to the mean to represent the typical criterion value of a cue sum category, because it provides a more robust measure of central tendency. However, the use of the median implies that in a learning situation in which the decision maker gets familiar with the estimation problem the criterion values of all encountered objects need to be stored to compute the median. In contrast, using the mean would not require storing all criterion values—the criterion value of each new object could be used to update the mean. More specifically the mean $M_{k,n}$ of all encountered objects n falling in the cue sum category k can be determined by $M_{k,n} = M_{k,n-1} + (1/n) \cdot (x_{k,n} - M_{k,n-1})$, where $x_{k,n}$ represents the criterion value of the newly encountered objects and $M_{k,n-1}$ represents the mean of all objects encountered before. Thus, this updating rule requires less demand on memory, because the decision maker only needs to store the mean and the number of objects encountered so far. In the reported studies we do not model the learning process of how people represent cue sum categories, but it is a task for future research to test whether the use of the mean as opposed to the median might have the advantage of providing a better description of the initial learning process.

2. In the case of binary cue information the multiplicative similarity rule of the original context model is a special case of a multidimensional scaling approach to modeling similarity as used by the generalized context model (Nosofsky, 1992). Thus the exemplar model we used is comparable to Nosofsky's model in how similarity is modeled.

3. According to Albers (2001), spontaneous numbers are multiples of powers of 10 $\{a\ 10^i: a \in \{1, 1.5, 2, 3, 5, 7\}\}$, where i is a natural number. They form a psychologically sensible set of coarse numbers, which increase in their crudeness as the numbers increase in magnitude (see also Hertwig et al., 1999).

4. The training and test sets in Studies 1 and 2 were selected on the basis of the predictions of the standard exemplar model. For the sake of clarity we focus throughout this article on the simplified exemplar model with one parameter—the strongest version of the exemplar model; however, the simplified versions of the models were only included post hoc. Thus the design of Studies 1 and 2 were based on the standard version of the exemplar model.

5. We used two measures of goodness-of-fit, the *RMSD* between the estimation and the criterion and the coefficient of determination (r^2). These two measures are closely related but capture slightly different aspects of the model fit. Both are based on the sum of squares error (*SSE*); but whereas the *RMSD* averages the *SSE* across the number of estimations, the coefficient of determination puts the squared error in relation to the total variance. This relationship can be demonstrated by the following equations:

$$RMSD = \sqrt{SSE(w_{LSE})/m},$$

$$r^2 = (1 - SSE(w_{LSE}))/SST$$

where *SSE* is the sum of squares error; w_{LSE} the parameter that minimizes *SSE* (w), *SST* the sum of squares total defined by $\sum_i (y_i - y_{mean})$, and m the sample size (cf., Myung, Pitt, & Kim, 2005, p. 426).

6. The cue–criterion correlations of some cues fluctuated around zero. For example, in the first three quarters of the training phase of the linear condition the third cue was positively correlated with the criterion, but in the last quarter of 55 trials it was negatively correlated with the criterion.

7. Additionally, we fitted the exemplar model in the exact same way as reported by Juslin et al. (in press) and replicated the reported fits. We chose an iterative fitting procedure to model the growing memory base during the training phase, because in Juslin et al. the

criterion values were not deterministic but changed for the identical profiles due to a random error. In Studies 1, 2, and 3 the iterative fitting procedure was unnecessary due to the deterministic criterion values.

8. Our results for the regression model differ from the results reported by Juslin et al. (in press) because we implemented an unconstrained regression model. Juslin et al. restricted the intercept to be the minimum criterion value in the training set and all cue weights had to add up to 1 (see Juslin et al. in press, Appendix, p. 49). The unconstrained regression model performed better in both conditions—in particular in the multiplicative condition our results were much better than those reported by Juslin et al.

Tables

Table 1

Mobile Phone Example for Illustrating the Predictions of the Models

	Phone A	Phone B	Phone C	Phone D	Phone Psi	Phone Omega
Cues						
Digital camera	-	-	-	+	+	+
Internet access	-	+	+	-	+	-
Weight	-	-	+	+	+	-
Display size	+	-	-	+	-	-
Criterion (selling price, in dollars)	10	20	30	100	?	?
Estimations of the models (in dollars)						
Mapping	15	15	30	100	100	15
Regression	10	20	30	100	110	90
QuickEst	15	15	20	50	30	15
Exemplar	10	20	30	100	30	43

Note. A plus sign indicates a positive cue value—for example, the phone possesses a digital camera or is lightweight; A minus sign indicates a negative cue value—for example, the phone does not possess a digital camera or it is heavy.

Table 2

Models' Average Accuracies (*Root Mean Square Error*) *in the Simulation Study for the Two Environments*

Model	J-shaped				Linear			
	Calibration sample		Validation sample		Calibration sample		Validation sample	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Mapping	14.3	3.5	15.3	1.6	21.6	5.1	25.9	6.4
Regression	14	2.4	16.5	1.2	20.9	4.7	27.7	6.3
QuickEst	14.8	1.7	14.9	1.1	24.8	3.5	28.3	3.5
Exemplar	12	3.5	15.8	1.7	17.5	4.9	27.2	6.2

Note. The models were initially fitted to the calibration sample, which contained 50% of the objects; the validation sample was used to cross-validate the results and comprised the other 50% of objects. Model predictions in the validation sample were made by using the parameter values derived in the calibration sample. The variation in model accuracy was higher in the linear environment, as the design in the linear environment varied over a higher number of correlations, and magnitude of correlation affected the accuracy of the models.

Table 3

Task Structure of Study 1

Exemplar no.	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5	J-shaped criterion	Linear criterion
1	0	0	0	0	0	20	20
2	0	0	0	1	0	23	60
3	0	0	0	0	0	26	80
4	0	1	0	1	0	28	140
5	0	0	0	1	0	29	160
6	0	0	1	0	1	33	220
7	0	1	0	1	0	34	240
8	0	0	1	0	1	35	260
9	0	1	0	0	0	40	300
10	0	1	0	1	1	41	420
11	0	0	0	1	0	47	440
12	0	1	1	0	1	52	480
13	0	1	0	1	0	62	540
14	0	1	0	1	1	71	640
15	0	1	0	0	0	110	660
16	0	1	0	1	1	160	720
17	1	1	1	1	1	200	840
18	0	1	0	1	1	250	880
19	1	1	1	1	1	500	920
20	1	1	1	1	1	1,000	1,000

Table 4

Correlations Between Cues and Criteria in Study 1

	Cue 1	Cue 2	Cue 3	Cue 4	Cue 5
J-shaped criterion	.79	.35	.48	.30	.42
Linear criterion	.65	.66	.37	.39	.62

Table 5

Models' Average Accuracies in Predicting Participants' Estimations in Study 1

	Linear environment				J-shaped environment			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
				Training set				
RMSD	149	93	168	138	125	98	125	116
SD	35	26	23	62	41	40	40	37
r^2	.75	.89	.69	.81	.77	.77	.76	.76
SD	0.16	0.07	0.10	0.08	0.17	0.17	0.18	0.17
				Test set				
RMSD	158	166	285	161	139	342	166	166
SD	49	56	46	40	93	124	101	70
r^2	.68	.67	.31	.67	.55	.20	.39	.47
SD	0.17	0.17	0.07	0.13	0.23	0.23	0.24	0.17

Note. The number of participants was 29 in the linear environment and 30 in the J-shaped environment. The exemplar model had one free parameter.

Table 6

Mean Consistency of the Participants in the Test Set of Study 2

	Linear				J-shaped			
	r	*SD*	*RMSD*	*SD*	*r*	*SD*	*RMSD*	*SD*
Old profiles	.89	0.08	129	48	.91	0.10	89	54
New profiles	.67	0.17	146	56	.78	0.17	86	42

Note. There were 25 participants in the linear environment and 25 in the J-shaped environment.

Table 7

Models' Average Accuracies in Predicting Participants' Estimations in the Test Phase of Study 2 (Test Set)

	Linear				J-shaped			
	Mapping	Regression	QuickEst	Exemplar	Mapping	Regression	QuickEst	Exemplar
Old								
RMSD	160	139	244	165	92	156	147	88
SD	35	36	33	35	26	9	24	31
r^2	.68	.76	.33	.68	.84	.54	.69	.85
SD	0.13	0.11	0.09	0.12	0.11	0.10	0.14	0.11
New								
RMSD	174	172	246	184	100	216	163	148
SD	43	58	51	42	58	34	33	24
r^2	.38	.50	.25	.37	.61	.44	.29	.50
SD	0.19	0.18	0.14	0.15	0.19	0.14	0.22	0.19
Total								
RMSD	167	154	246	174	99	186	156	118
SD	34	44	35	32	13	17	21	18
r^2	.60	.67	.27	.58	.77	.36	.44	.70
SD	0.13	0.14	0.09	0.15	0.13	0.08	0.11	0.09

Note. There were 25 participants in the linear environment and 25 in the J-shaped environment.

Figure Captions

Figure 1.

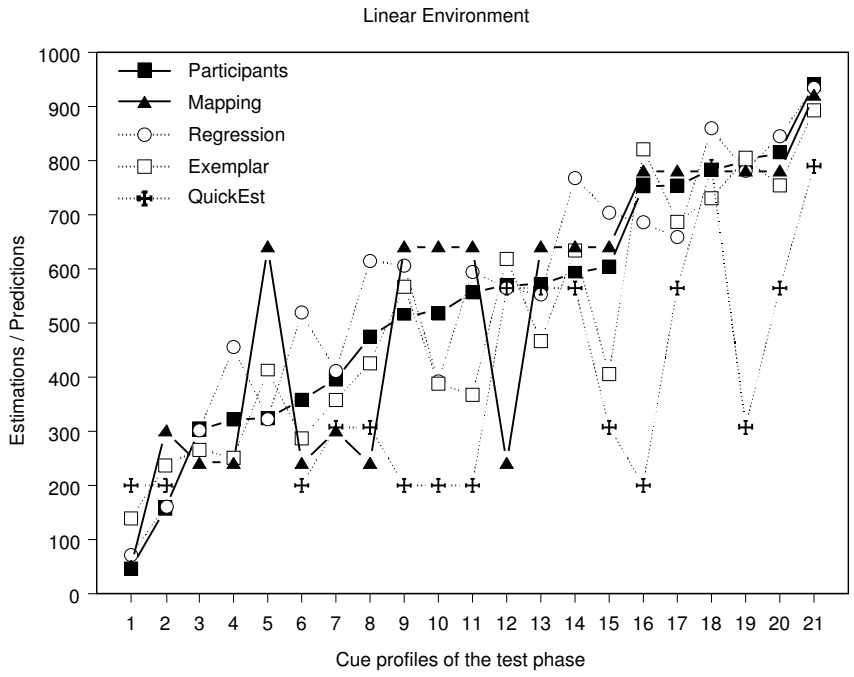
Models' predictions and participants' estimations in the test phase for (A) the linear environment and (B) the J-shaped environment of Study 1. The profiles in the test set are rank ordered according to the participants' average estimations. In the linear environment, profiles 1, 2, 3, 5, 7, 12, 13, and 21 were included in the test and training set. In the J-shaped environment, profiles 1, 2, 3, 4, 5, 7, 8, and 21 were included in the test set and the training set.

Figure 2.

Models' predictive accuracies for the new profiles of the test phase of Study 3. The average root mean square deviation (RMSD) between the models' predictions and the participants' estimations for the linear and the multiplicative condition is depicted. The error bars represent the 95% confidence intervals.

Figure 1.

A


Linear Environment
Participants
Mapping
Regression
Exemplar
QuickEst
Estimations / Predictions
Cue profiles of the test phase

B

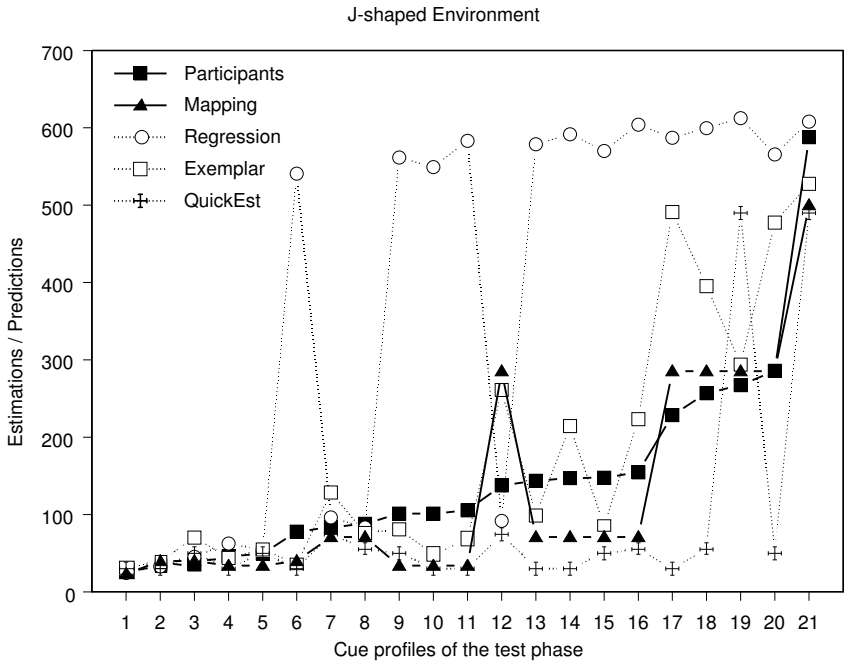
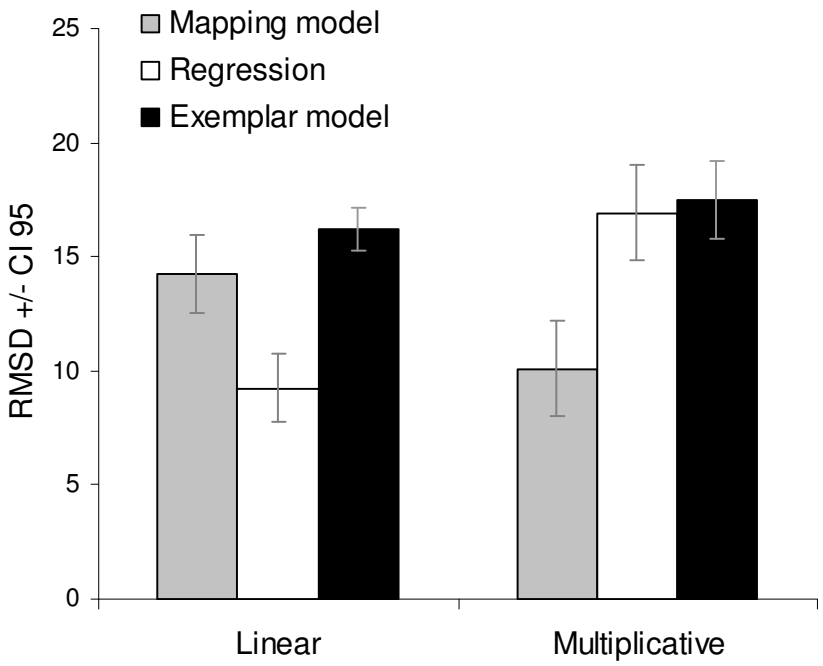

J-shaped Environment
Participants
Mapping
Regression
Exemplar
QuickEst
Estimations / Predictions
Cue profiles of the test phase

Figure 2.


Mapping model
Regression
Exemplar model
RMSD +/- CI 95
25
20
15
10
5
0
Linear
Multiplicative