

# Trade-offs in the Dark: Exemplar-Based Learning and Extrapolated Preference Functions

*Yanliu Huang*

*Marketing Department, The Wharton School*

*Under the supervision of Professor Robert Meyer*

Many choices require us extrapolate preferences formed in one domain of experience to an unfamiliar new one. An Ohioan who has extensive experience choosing among apartments in Toledo may, for example, find this experience to be of limited help when choosing an apartment in New York or London. How do individuals make choices when asked to extrapolating existing knowledge to new domains? The answer to this question, perhaps surprisingly, is far from fully known. While there has been considerable work investigating such related problems as inference formation and behavior in prediction tasks (e.g., DeLosh, Busemeyer, and McDaniel 1997), less is known about either the process individuals use to make trade-offs among unfamiliar ranges of attributes or the algebraic structure of such trade-offs.

Central to this research is a hypothesis that judgments made in novel settings are dominated by exemplar-based processes (such as pattern-matching) that associate predicted utilities for new stimuli based on their similarities to familiar stimuli (e.g., Juslin, Olsson, and Olsson 2003; Shanks and Darby 1998). We show that for a general class of pattern-matching algorithms extrapolated preference functions revealed by individuals using such processes will exhibit two central properties:

1. Exaggerated concavity, with individuals increasingly under-estimating the likely hedonic impact of increasingly extreme attribute values (akin an anchoring bias); and
2. Asymmetry in this bias, with the implicit anchoring bias being more extreme when individuals are asked to extrapolate to inferior (less preferred) values of an attribute than superior values.

We tested these predictions using data from an experiment in apartment learning. Participants were asked to act as agents to learn to predict a target customer's apartment preferences based on four attributes: travel time to work, apartment appearance, security of apartment location, and rent. The target customer's preferences were given by a linear-additive deterministic multi-attribute preference function unknown to the participant.

Each attribute had three levels, and a fractional design (i.e., 12 combinations) allowed orthogonal estimation of all main effects of the four attributes. Participants were instructed to first make a prediction of the target customer's preference for each of the 12 apartments on a 90-point scale, and then the true preference was revealed to them after each prediction. After 12 training instances were judged, participants predicted the target customer's preferences for a new set of 12 apartments without feedback. At this time they were randomly assigned to one of three conditions: a set of 12 better profiles with greater values on each attribute, a set of 12 poorer profiles with worse values on each attribute, and a set of 12 similar profiles.

Our hypothesis was that when participants were asked to predict preferences for apartments for new, extreme, attribute values implicit predictions would display a asymmetric anchoring effect, increasingly underestimating the true marginal utility for increasingly negative attribute values and but only slightly under-estimating the true marginal utility for increasingly positive attribute values. The data strongly supported this prediction. As one example, FIGURE 1 plots the true versus estimated marginal means for the “travel time” attribute where the training phase focused on the attribute levels of 15, 20, and 25 minutes and participants performed better on the same arrange of attribute levels and the better level ranges of 5, 10, and 15 minutes than on the worse level ranges of 25, 30, and 35 minutes.

FIGURE 1 True and estimated marginal means for each level of "Travel Time" attribute

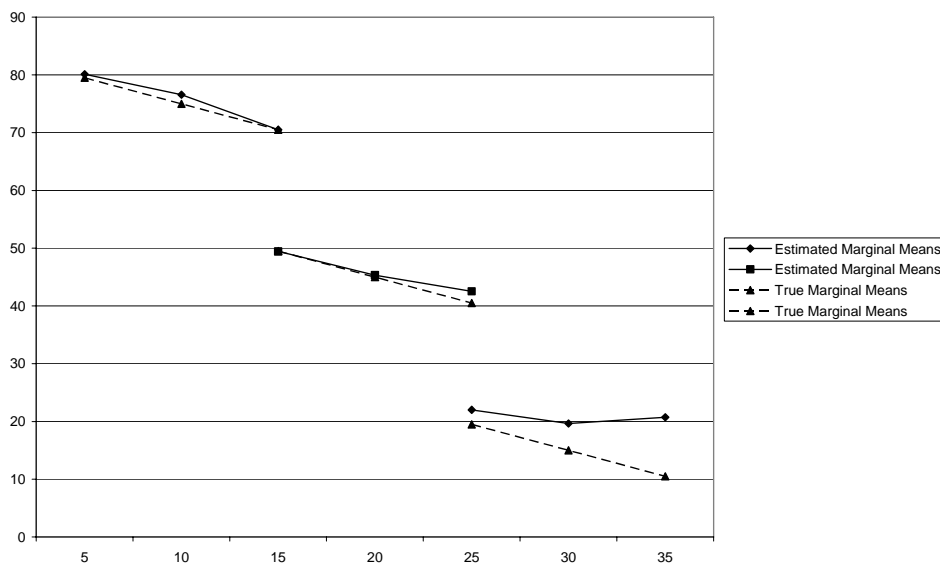


Figure 1

Further reinforcing this effect, over all judgments participants’ ability to extrapolate was decidedly asymmetric: the correlation between estimated and true preferences were 0.58 and 0.66 respectively when the new apartments took on attribute values that were similar to or better than those in the training set, respectively, but only 0.33 when the attribute values were worse condition.

While the data are consistent with the judgment pattern that would be predicted if participants had much better developed exemplars for “good” apartments than “bad” apartments, the fact that no apparent anchoring effect was observed for positive extrapolations argues against a low-level pattern-matching process a singular description of the data. Specifically, consistent with the findings of DeLosh, Busemeyer, and McDaniel, 1997, given positive attribute values participants acted as if they were making use of higher-order skills of linear extrapolation. A challenge to future work will be to develop a theoretical framework that makes testable predictions about when higher- versus lower-order processes in extrapolation are invoked.