# Past Experience Influences Judgment of Pain: Prediction of Sequential Dependencies

**Benjamin V Link (link@colorado.edu)**
Dept. of Computer Science, University of Colorado at Boulder
Boulder, CO 80309 USA

**Brittany Kos (brittany.kos@colorado.edu)**
Dept. of Computer Science, University of Colorado at Boulder
Boulder, CO 80309 USA

**Tor D. Wager (tor.wager@colorado.edu)**
Dept. of Psychology and Institute of Cognitive Science, University of Colorado at Boulder
Boulder, CO 80309 USA

**Michael Mozer (mozer@colorado.edu)**
Dept. of Computer Science and Institute of Cognitive Science, University of Colorado at Boulder
Boulder, CO 80309 USA

## Abstract

Recent experience can influence judgments in a wide range of tasks, from reporting physical properties of stimuli to grading papers to evaluating movies. In this work, we analyze data from a task involving a series of judgments of pain (discomfort) made by participants who were asked to place their hands in a bowl of water of varying temperature. Although trials in this task were separated by a minute in order to avoid sequential dependencies, we nonetheless find that responses are reliably influenced by the recent trial history. We explore a space of statistical models to predict sequential dependencies, and show that a nonlinear autoregression using neural networks is able to predict over 6% of the response variability unrelated to the stimulus itself. We discuss the possibility of using decontamination procedures to remove this variability and thereby obtain more meaningful ratings from individuals.

**Keywords:** Sequential Dependencies; Judgment Models

## Introduction

When asked to make absolute judgments in an experimental setting individuals use *anchoring* or *primacy*: information presented earlier in time serves as a basis for making judgments later in time (Tversky & Kahneman, 1974). The need for anchors is due to the fact that individuals are poor at or possibly incapable of making absolute judgments and instead must rely on reference points to make relative judgments (Laming, 1984 ; Parducci, 1965 ; Stewart, Brown, & Chater, 2005). The literature in experimental and theoretical psychology exploring *sequential dependencies* suggests that reference points change from one judgment or rating to the next in a systematic manner.

Teachers are cognizant of potential drift when grading papers and the necessity of comparing early papers to those graded later. Sequential dependencies arise in a myriad of common tasks, such as responding to surveys, questionnaires, and evaluations. A relatively unexplored field of sequential effects involves online recommendation engines. Netflix, Amazon, and Google consistently recommend products through advertisements that they think you would be interested in buying. Could these recommendation engines be improved by observing how you are rating products sequentially? By mitigating the influence of recent judgments, recommendation engines could make more meaningful and accurate predictions for what products you are interested in. Even small improvements in these engines can mean large income increases. By having the best recommendation engine you not only sell more products, but you draw more users.

Carefully controlled laboratory studies of sequential dependencies, dating from the 1950's (Miller, 1956), consist of rating unidimensional stimuli, such as the decibel level of a tone, or the length of a line. These studies suggest that across many such domains, responses convey not much more than two bits of mutual information with the stimulus (Stewart et al., 2005). Various types of judgment tasks have been studied including *absolute identification*, where the individual's task is to specify the value of the stimulus level (e.g., 10 levels of loudness), *magnitude estimation*, where the task is to estimate the magnitude of the stimulus which could vary continuously along a dimension, and *categorization*, where the task requires the individual to label stimuli by range. Due to the large size of responses in absolute identification and categorization tasks, and because individuals aren't usually aware of the discrete stimuli in absolute identification tasks, there isn't a qualitative difference among tasks. Typically, feedback is provided in absolute identification and categorization tasks. Without this feedback, explicit anchors against which stimuli can be assessed wouldn't exist.

The consequences of sequential effects can be complex. Normally, on trial $t$ of an experiment, trial $t-1$ has the largest influence on ratings and earlier trials—$t-2$, $t-3$, and so forth—have successively diminishing influences. Both the stimulus and response on a previous trial can have an effect, which makes sense if individuals formulate a response to the

current stimulus by analogy to the relationship between previous stimuli and responses. Two types of effects are observed: an *assimilative* effect occurs when the current response moves in the direction of stimulus or response from a previous trial; a *contrastive* effect is one that moves away. Analyzing recency effects using assimilation and contrast is complex and theory dependent (DeCarlo & Cross, 1990).

Because cognitive scientists are aware the recent trial history can influence responses to a stimulus, studies are often designed to limit or completely avoid sequential dependencies. Increasing the number of response categories and varying the type and frequency of anchors are common methods to mitigate sequential dependencies (Mumma & Wilson, 2006 ; Wedell, Parducci, & Lane, 1990). Another possible approach is to increase the intertrial interval, on the assumption that recency effects decay to some extent with the passage of time. In this paper we will describe data that was collected in which trials were separated by sixty seconds, in the hope that sequential effects would be suppressed. We show that even in this scenario, significant sequential effects do occur. Fortunately, we also show that they can be predicted and there is therefore hope for removing the contaminative effect they have.

## Experimental Data

The data we analyze in this paper come from experiments conducted in Tor Wager's lab at Columbia University over a period of several years. Wager studies brain activity associated with pain and placebo effects. Participants are asked to judge the level of discomfort (pain) associated with pools of water varying from 32° to 53° Celsius, with the higher temperatures associated with more discomfort. Each participant in an fMRI study begins with a calibration procedure that attempts to determine the mapping between water temperature in degrees Celsius and pain rating using a 10 point rating scale, 1 being lowest level of pain, and 10 being the highest.

The calibration procedure involves 24 trials, the goal of which is to determine temperatures that correspond to subjective pain levels 2, 5, and 8 on a 10-point scale. This goal is achieved by an adaptive algorithm that explores the range of temperatures in order to obtain data that is well fit by an affine transformation from temperature to pain level via least squares regression. Consequently, the order of stimuli is not entirely random, because the temperature is chosen on a trial to provide the most information about the transformation. However, because the procedure jumps pseudo-randomly between calibration of low, medium, and high pain levels, there is significant trial-to-trial variability in the temperatures. From the participants' perspective, there is no trial-to-trial predictability of temperature, and the temperature levels fluctuate without any perceptible pattern.

We obtained pain judgment data from a total of 284 participants. Although the participants were part of 17 distinct experiments, the calibration procedure was identical in all experiments.

## Analysis of Pain Judgment Data

Our first goal is to determine whether sequential dependencies are present in the data. One intuitive approach is simply to plot the response to the current stimulus as a function of the previous stimulus. Because of the sparsity of data, the closest we could come to making such a graph is to partition the stimuli into five ranges, and plot—for each stimulus partition—the response as a function of the previous stimulus partition, as is shown in Figure 1. Each point on the graph is an expectation over all trials of all participants who were shown a particular stimulus on trial $t$, $S(t)$, following a previous stimulus, $S(t-1)$; this response is denoted $E[R(t)|S(t),S(t-1)]$. Because we are concerned with how responses deviate based on earlier trials, we subtract out the mean response to the current stimulus, $E[R(t)|S(t)]$.

If previous trials had no influence, each curve in the Figure would be flat, indicating that the mean-subtracted response on trial $t$—depicted on the ordinate—is independent of the previous stimulus, $S(t-1)$—depicted along the abscissa. However, the pattern we observe is quite different. Four of the five stimulus partitions show a clear negative slope: the response to the current stimulus tends to decrease as the previous stimulus increases. This negative slope is a contrast effect. A low value of $S(t-1)$ tends to cause $S(t)$ to be given a higher rating, and a high value of $S(t-1)$ tends to cause $S(t)$ to be given a lower rating.

The fifth partition of $S(t)$ in Figure 1—reflecting the temperature range 32–37°, seems to be relatively unaffected by the previous trial. It is quite common for the extreme stimulus values to be less influenced by recency that the intermediate stimulus values, due to the fact that the extreme stimuli become effective anchors. For example, (Mozer et al., 2010) found very weak sequential effects for the extrema in a line length judgment task.

The sequential effects can be quite substantial. For the 43.5–45° range, the response fluctuated by 4 points on the 10 point scale due to the previous stimulus.

In Figure 1, we partitioned the stimulus range in order to obtain roughly equal numbers of judgments in each partition. We explored several other partitioning schemes—including selecting equal temperature bin widths and bin widths that yielded an equal range over responses— and all produced graphs qualitatively similar to Figure 1.

Although the graph strongly suggests the existence of sequential effects in the pain judgment data, one must interpret it with caution because the data points represent averages over many individuals and many trials. It's altogether possible that even if sequential effects are robust and measurable for aggregated data, it will be impossible to detect them for a particular individual on a particular trial. If our long-term goal is to obtain more meaningful ratings from individuals by removing the contamination from recent trials, then we need to show that it is possible to account for variability in an individual

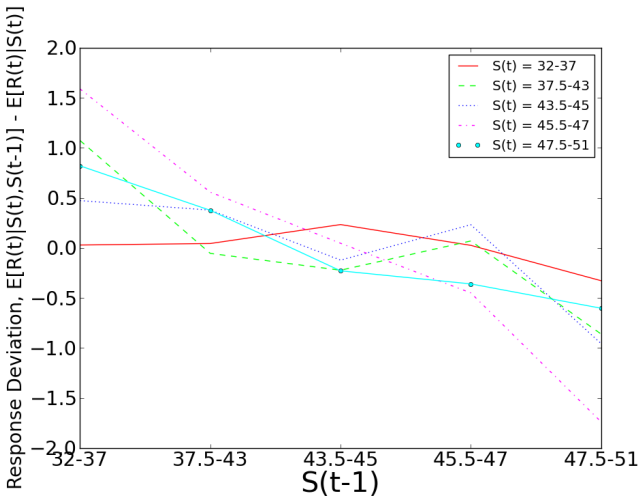


Figure 1: Each curve represents the average response deviation for a given range of the current stimulus, $S(t)$, as a function of the previous stimulus, $S(t-1)$. The response deviation specifies how much the expected response differs from the overall mean response. Each data point is an average over many trials and many participants.

trial based on recent history. In past work, we (Wilder, Jones, & Mozer, 2009) found that sequential effects could explain upward of 95% of variability in *aggregated* responses on a very simple two-alternative forced choice task but only about 25% of variability in individual trials.

Thus, our next goal is to show that we can reliably detect sequential effects on an individual trial in our data set. We approach this goal by constructing mathematical models that describe how recent history—e.g., $S(t-1)$, $R(t-1)$, $S(t-2)$, and $R(t-2)$—influences the current response, $R(t)$. There is a rich psychological modeling literature that attempts to explain sequential effects in judgment, absolute identification, and choice tasks. DeCarlo et Cross (1990) describe a thirty year history of models that all characterize the current response as a linear function of the previous stimulus and/or response. Other models are in the same form (e.g., Stewart et al., 2005 ; Wilder et al., 2009), although they include stimuli and responses from two and more trials back in the linear model. For judgment of physical magnitudes (e.g., pitch), the simple linear form of the models is obtained by log transforming the raw stimulus intensities. The primary distinction among the various linear models is the coefficients that weight terms in the model, and constraints assumed to operate among these coefficients. To represent this large class of models, we explore linear predictive models and treat the coefficients as free parameters that are fit to the data.

In the literature, a class of psychological models assume that past trials provide reference or *anchor* points relative to which the current trial is compared (e.g., Parducci, 1965 ; Petrov & Anderson, 2005). One key feature of these anchors is that generalization from the anchors to the current trial is similarity dependent (Petrov & Anderson, 2005). To allow for nonlinear effects such as similarity dependence, we also consider a class of models that is primarily linear but allows some degree of nonlinearity, specifically via the computation of distances between the current and previous stimuli.

The models we explore predict the response on the current trial given recent trial history, and we attempt to show that these models outperform a baseline model that predicts based solely on the current stimulus. We begin by describing the baseline model.

## Baseline Regression

We assume that individuals map the stimulus continuum to the response continuum using an affine transformation, and thus we can predict an individual's response as

$$\hat{R}(t) = \beta_0 + \beta_1 S(t), \tag{1}$$

where the coefficients $\beta = \{\beta_0, \beta_1\}$ may differ from one individual to the next. Although Weber's law suggests that transforms from physical stimulus magnitudes to internal representations should be logarithmic, an inspection of the data reveals a roughly linear relationship, as depicted in Figure 2 for six different participants. The red circles indicate responses on individual trials. The solid green line represents the least squares regression, which obtains the coefficients $\beta$ and the blue squares represent the improved fit of a model that we have yet to describe.

The residual error, $\rho(t) = R(t) - \hat{R}(t)$, might simply be due to factors outside of the experimental context, such as the individual's attentional state, or the residual error might be attributable to some systematic influence, such as sequential dependencies in formulating a response. We will investigate this latter possibility via computational models. We build several types of models to predict the residual error. If the recent trial history helps to reduce the residuals, we have evidence for sequential dependencies in this experimental study.

Although we obtain $\beta$ coefficients for each individual separately, we build a single sequential-dependency model for all individuals. The reason for this decision is that we have relatively sparse data from each individual—a total of 24 trials—and some of the sequential-dependency models we consider have a large number of free parameters, and can only be constrained with large amounts of data. However, if we do find significant variability that can be explained *across participants* from a model of sequential dependencies, the explanatory power of a model tailored to an individual is potentially even greater.

We define the baseline fit via the root mean squared error,

$$RMSE_{baseline} = \left( \sum_i \sum_t \rho_i(t)^2 \right)^{\frac{1}{2}}, \tag{2}$$

where $i$ is an index over participants, $t$ is an index over trials, and $\rho_i(t)$ denotes the residual from the regression for participant $i$ on trial $t$. Intuitively, the RMSE indicates how large

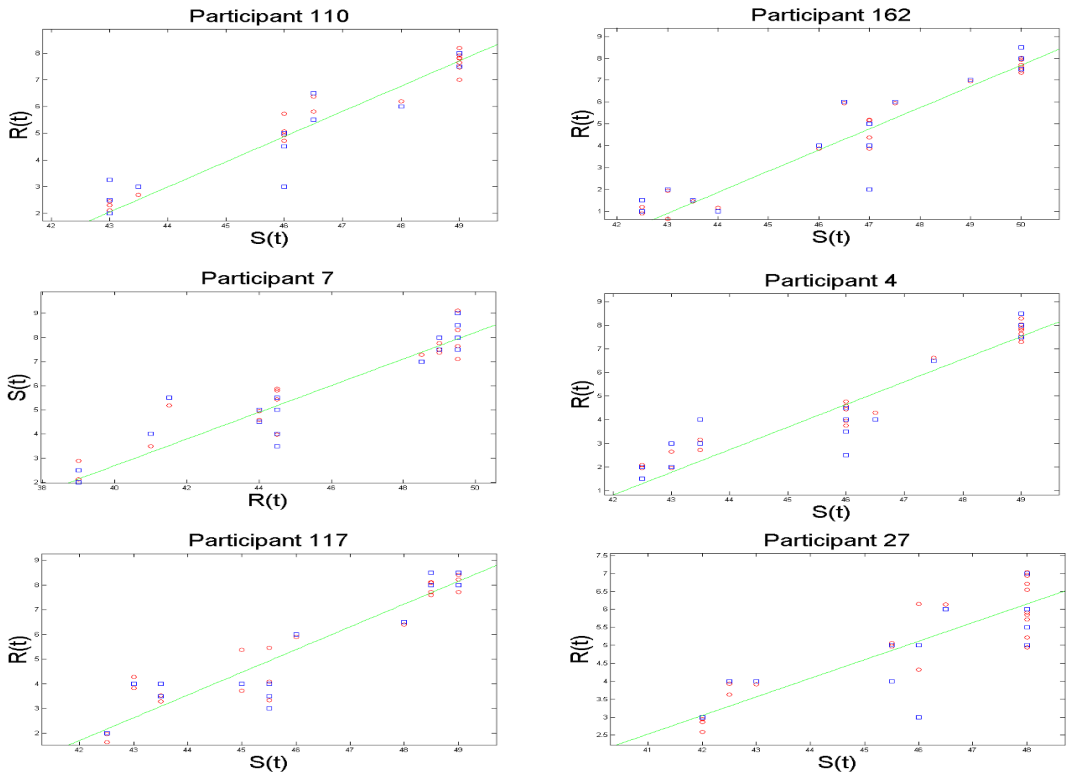


Figure 2: Pain judgment data from six participants. Each red circle in the scatterplot represents a single stimulus-response pair, where the stimulus level is depicted along the abscissa and the response level is shown on the ordinate. The solid green line represents the least square regression. The blue squares show the cross validated prediction of the best model we explored.

a deviation a model produces from the actual response an individual makes. In our data set, $RMSE_{baseline} = 1.2502$, indicating that the baseline model produces a typical deviation slightly larger than one unit on the 1-10 response scale. We will evaluate all sequential-dependency models in terms of how effectively they reduce $RMSE_{baseline}$.

We use cross validation—the standard paradigm from machine learning and statistics—to estimate the effectiveness of a model. In all simulation results reported below, we perform 10-fold cross validation on our set of participants, using data from 9/10th of the individuals for training and then hold out 1/10th for evaluation, and repeating the validation step for each of 10 hold out sets.

## Models

In this section, we describe a series of models that are designed to predict the residuals from the baseline model, i.e., to predict the structure in the data due to the sequence and unrelated to the current stimulus. If the model has no predictive ability—i.e., it predicts 0 for each residual—it will perform no better than the baseline model. If the model is able to predict all of the residual, the RMSE will drop to 0. Thus, the models we explore should yield RMSE values between 0 and $RMSE_{baseline}$.

We explored a space consisting of eight distinct models which differ along three binary dimensions. The dimensions of the model space are motivated by existing theories of sequential dependencies. We now describe the three dimensions of our model space: the model *class*, *history*, and *order*.

**Model Class: Regression Versus Neural Net.** Most models of sequential effects assume some linear influence of previous trials and some nonlinear influence. Thus, we consider both linear and nonlinear regression. We use a three-layer back propagation neural network as a generic nonlinear regression model. All neural nets had 10 hidden units, used a tan-sigmoid transfer function for the hidden layer, a linear transfer function for the output layer, and were trained with early stopping. The early stopping procedure reserves 10% of the training data for validation, and terminates training when the error rate on the validation set begins to rise. (The training and validation sets are distinct from the cross-validation hold-out set used to evaluate the model.) We experimented with networks of different sizes and the results were comparable to what we present below.

**Model History: One Versus Two Trials Back.** All theories of sequential effects assume a diminishing influence of more distant trial history, usually with an exponential fall off. Many models consider only the previous trial, but gener-

ally modelers find a benefit of including longer histories. We explored what we will term *one-back* and *two-back* models. One-back models utilized the previous stimulus and response, $S(t-1)$ and $R(t-1)$. Two-back models utilized the previous two trials, $S(t-1)$, $R(t-1)$, $S(t-2)$, and $R(t-2)$.

**Model Order: First Versus Second.** Some models of sequential effects suppose that the spillover from trial $t-n$ to trial $t$ is dependent on the similarity of the stimuli on trials $t-n$ and $t$ (DeCarlo & Cross, 1990 ; Petrov & Anderson, 2005). Given that the stimuli in our data were temperature levels from a continuous scalar dimension, the similarity can be measured in terms of the squared Euclidean distance, $(S(t)-S(t-n))^2$. To allow models to readily utilize this measure, we included as model regressors the terms $S(t)^2$, $S(t)S(t-n)$, and $S(t-n)^2$ for a model that considers the $n$-back trial. With these three additional regressors, it is a linear operation to compute squared distance.

## Simulation Results

The three binary dimensions of our model space specify eight distinct models. We trained each model to predict the residuals of the baseline model, $\rho(t)$, across all individuals in the data set. Table 1 shows the RMSE for the different models obtained via cross validation, and also the percentage improvement of the model over the baseline. This latter quantity is simply

$$\%Improvement = 100(RMSE_{base} - RMSE_{model})/RMSE_{base}$$

As we anticipated, the nonlinear regression model—the neural net—performed far better than the linear model. Indeed, we find little leverage from the models that are purely linear in $S(t-n)$ and $R(t-n)$. The nonlinearity of the second order models—which include terms quadratic in $S(t)$ and $S(t-n)$—also apppears to have improved prediction significantly.

Finally, the two-back models performed better than the one-back models. The boost provided by trial $t-2$ is generally smaller than the boost provided by trial $t-1$, consistent with the exponential decay of influence of previous trials found empirically in the sequential effects literature.

Figure 3 shows the RMSE represented as a bar graph with standard errors indicating the uncertainty in the RMSE across cross-validation splits of the data. Inspecting Figure 3, one surprising finding is that the neural net yields not only larger improvements in RMSE, but also highly consistent improvements: the standard error in the RMSE estimate is quite small.

The most complex model—the second order neural network model with two-back history—is evidently the best. This model produces a more than 6% reduction in error over the baseline model. That is, the sequential influence of previous trials on judgment explains 6% of what appears to be noise in the data. This result is all the more impressive considering that a single model is constructed for all participants, and there may well be significant individual differences in the nature of sequential effects.

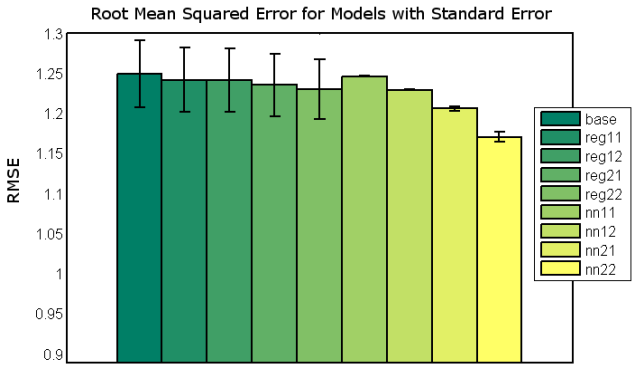


Figure 3: Root Mean Squared Error (RMSE) for the eight models. The errorbars indicate +/- one standard error of the mean. base=baseline model; reg*xy* = *x*-order *y*-back linear regression; nn*xy* = *x*-order *y*-back neural network

Figure 2 shows some examples of data and the corresponding model fit. Each graph represents a different individual. Each red circle plots the response produced by the individual (on the ordinate) to a stimulus (on the abscissa). The solid green line is the best fitting linear regression, and $\rho$ is the deviation from the red circles to the green line. The blue squares show the predictions of the second-order neural net with two-back history. (For this simulation, the neural net was trained on data excluding the individual whose data on which predictions are plotted. Thus, the red squares are not fits to data, but predictions from a pretrained model.) The prediction of the model is an improvement over the baseline if the red circle is closer to the corresponding blue square than to the green line. For most trials, the Figure shows that a better prediction of the response is made by considering the influence of recent trial history (the blue squares) than by using the current stimulus alone (the green line).

## Conclusion

Through our simulation models, we find that sequential dependencies can explain more than 6% of the 'noise' in judgments of pain. To gauge what 6% means, consider that the much-publicized Netflix competition aimed to improve predictions of movie ratings by 10% (Koren, August 2009). The winners of the competition used many different methods to reach this goal, most of which produced a much smaller improvement than 6%. Sequential dependencies likely played a role in the Netflix data, given that individuals often rate movies in consecutive bursts.

The 6% improvement is particularly interesting given that our data come from an experiment that was designed to avoid sequential dependencies by spacing judgements a minute apart. It seems likely that the effects would have been larger in magnitude if judgments had been more closely spaced in time.

Sequential dependencies are ubiquitous in cognitive tasks.

Table 1: RMSE Results for Sequential-Dependency Models

| Model Class | Model Order | Model History ($n$ back) | Cross-validation RMSE | % Improvement Over Baseline |
|---|---|---|---|---|
| Regression | 1st | 1 | 1.2423 | 0.63% |
| | | 2 | 1.2418 | 0.67% |
| | 2nd | 1 | 1.2360 | 1.14% |
| | | 2 | 1.2301 | 1.61% |
| Neural Net | 1st | 1 | 1.2469 | 0.26% |
| | | 2 | 1.2298 | 1.63% |
| | 2nd | 1 | 1.2064 | 3.50% |
| | | 2 | 1.1712 | 6.32% |

It's impossible to find a domain where sequential dependencies don't arise, from the simplest of choice tasks, to language use, to the control of attention (Mozer, Kinoshita, & Shettel, 2007). Cognitive scientists well appreciate that experimental design needs to take into consideration the possibility of sequential dependencies. Despite attempts to control for sequential dependencies, for example by increasing the intertrial lag or by requesting a judgment of the same item in many different contexts, sequential dependencies still inject a source of uncontrolled variability into human performance. Rather than attempting to mitigate sequential dependencies in the experimental design, perhaps it is more productive to design experiments that enhance sequential effects, because doing so will make the modeling of these effects easier and when sequential effects are large, other forms of response variability may be suppressed.

Having constructed quantitative models to predict sequential dependencies, there is hope of exploiting the same models to remove their influence. We have taken steps in this direction with our attempt to invert models such as those we presented in this paper to *decontaminate* judgments, and effectively remove the contribution of recent trials to responses (Mozer et al., 2010). Although we have been successful in decontaminating responses in a simple visual judgment task, extending the technique to more complex, naturalistic tasks requires better models of the contamination process by which previous trials affects current judgements. The work described in this paper suggests the importance of nonlinearity in modeling the influence of recent trials on behavior.

## Acknowledgements

The authors thank Natalie Johnson and Tal Yarkoni for assembling the pain judgment data used in our study. This research was supported by NSF grants BCS-0339103, BCS-720375, SBE-0518699, and SBE-0542013 (Temporal Dynamics of Learning Center).

## Références

DeCarlo, L., & Cross, D. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General*, *119*, 375–396.

Koren, Y. (August 2009). *The bellkor solution to the netflix grand prize.*

Laming, D. (1984). The relativty of "absolute" judgments. *Journal of Mathematical and Statistical Psychology*, *37*, 152–183.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review*, *63*, 81–97.

Mozer, M. C., Kinoshita, S., & Shettel, M. (2007). Sequential dependencies offer insight into cognitive control. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 180–193). Oxford, UK : Oxford University Press.

Mozer, M. C., Pashler, H., Wilder, M., Lindsey, R., Jones, M., & Jones, M. (2010). Improving human judgments by decontaminating sequential dependencies. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 1705–1713). La Jolla, CA : NIPS Foundation.

Mumma, G., & Wilson, S. (2006). Procedural debiasing of primacy/anchoring effects in clinical-like judgments. *Journal of Clinical Psychology*, *51*, 841–853.

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

Petrov, A., & Anderson, J. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review*, *112*, 383–416.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Wedell, D., Parducci, A., & Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors. *Personality and Social Psychology*, *58*, 319–329.

Wilder, M., Jones, M., & Mozer, M. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 2053–2061). La Jolla, CA : NIPS Foundation.