Individual Differences in Causal Learning

Laila Johnston (lailacjohnston@gmail.com)
Department of Mathematics, University of Central Florida
Orlando, FL 32816 USA

Noah Hillman (hillma1@stolaf.edu)
Department of Mathematics, Statistics, and Computer Science, St. Olaf College
Northfield, MN 55057 USA

David Danks (ddanks@cmu.edu)
Departments of Philosophy and Psychology, Carnegie Mellon University, Baker Hall 161
Pittsburgh, PA 15213 USA

Abstract

Causal inference from observed cases is a central cognitive challenge. There has been some evidence for individual differences in causal learning strategies, but prior work has not examined fine-grained sequences of judgments. In this paper, we report a large-scale model-fitting effort to determine the best-fitting causal inference models for individual participants. We fit a range of different model-types against multiple judgment sequences from each participant, thereby enabling comparisons of learning strategy both between- and within-participant. The model-fitting effort revealed some diversity in learning strategy along both dimensions, though individuals did exhibit some stability. Overall, however, the model fits were worse than expected, particularly when compared to the high accuracy reported for many of the models when used to predict group-level causal judgments. These results thus call into question whether these models might accurately describe the average behavior without accurately describing many (or any) individual's behaviors.

Keywords: causal learning; individual differences; associationist learning; Bayesian model

Introduction

Causal knowledge is central to many different cognitive activities. We use our understanding of the causal structure of the world to guide prediction, explanation, reasoning, decision, and control. Moreover, for many of these processes, causal knowledge—rather than simple observation—is necessary to avoid making numerous and/or significant errors. If I observe that someone has a symptom, then I can predict that they have a disease. But knowledge of the symptom-disease association is relatively uninformative about potential treatments, possible explanations of those observations, or many other important cognitive tasks.

There has unsurprisingly been a large body of cognitive science research on various aspects of causal cognition, including perception (e.g., Michotte, 1946/1963; Scholl & Tremoulet, 2000), inference (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011; Lu, Rojas, Beckers, & Yuille, 2016), and reasoning (e.g., Rehder & Burnett, 2005; Rottman & Hastie, 2013). In the present paper, we focus on how people learn the strength of a potential cause to bring about an effect. For example, what is the strength of this plant to produce a rash in particular individuals?

Much research on this question has focused on group-level patterns in causal cognition (though see, e.g., Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). For example, experiments and models of causal learning from sequences of cases usually aim to measure, predict, and explain the mean judgment across the full set of participants. This emphasis on group-level patterns has some natural intuitive appeal: if causal cognition really is central to successful functioning, then it should arguably not exhibit significant variation between people. And if the processes underlying causal cognition are (approximately) species-universal, then mean judgments or learning curves will remove the inevitable experimental noise to reveal those processes.

While this argument has some intuitive appeal, a closer examination of the experimental data suggests that the conclusion does not hold. For example, Figure 1 shows learning curves for twelve (not randomly selected) individual participants in Danks & Schwartz (2006).

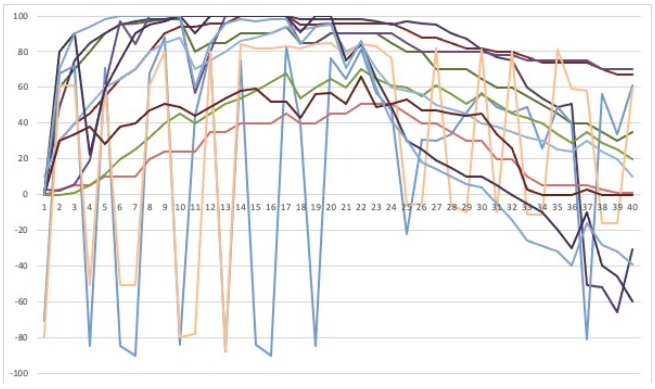

Figure 1: Example learning trajectories

These participants were all in the same experimental condition, and so all saw exactly the same sequence of cases. During that sequence, they were each asked after each case to judge the causal strength of a potential cause, given

everything that they had observed to that point. But although the participants in Figure 1 saw identical sequences of cases, their responses are quite different, not only in the particular numeric ratings, but also in the shapes of their learning curves. These participants certainly do not appear to be learning the causal strength using the same processes, though that appearance is potentially misleading.

There have been limited previous efforts to examine individual differences in causal learning. Steyvers, et al. (2003) developed a Bayesian model of causal structure learning from a sequence of cases, including parameters for choice optimality and individual memory. They then showed that participants naturally divided into several clusters, where they required all participants within a cluster to have the same parameter values. Danks & Schwartz (2005, 2006) examined individual differences in responses to surprising data or changes in the underlying causal structure, though principally through descriptive analyses of observed changes, rather than examination of the underlying cognitive processes. Their results suggest that people might be learning in qualitatively different ways, but do not resolve which of the many proposed theories are best for different individuals. Relatedly, a meta-analysis by Perales & Shanks (2007) suggested that no single model provides a best fit for average responses, which is exactly what would be expected if there are significant individual differences in learning strategy (see also Hattori & Oaksford, 2007; Lee & Lovibond, 2020).

In this paper, we directly examine individual differences in cognitive processes for causal learning. Many individual difference analyses (in other areas of cognitive science) posit a single process with one or more free parameters (e.g., working memory capacity). In contrast, we build upon the diverse set of potential theories previously proposed in the literature, and also ones developed for this paper. We fit those models to the individual-level data in both of Danks & Schwartz (2005, 2006). In particular, we consider both the best-fitting model for each individual participant in each condition, and also the best-fitting single model for each participant across all conditions. These different analyses enable us to explore both between- and within-individual differences in causal learning strategy. The next Section explains the theories that we considered. We then describe the specific experimental data and model-fitting procedure that we used, before turning to the analyses and remaining open questions.

Competing Models

This analysis considered fifteen causal learning models: seven models from the existing literature; seven novel models developed specifically for this analysis; and a "null model" (not further discussed) that simply predicted zero for every causal judgment ("Zero Correlation"). The seven previously-proposed causal learning models were selected based on three criteria. First, the selected models were all computationally fully-specified, so quantitative causal strength predictions could readily be generated. Second, the models all provide predictions on a case-by-case basis, rather than being limited to only long-run predictions. Third, the selected models have each been developed and tested in multiple experimental papers. Importantly, the existing models were originally proposed with differing interpretations; some were intended as mechanistic models, while others were presented at the computational level (Marr, 1982) that does not carry mechanistic commitments. For the purposes of this analysis, we remain agnostic about the commitments of any particular theory, and consider them as computational proposals.

In this section, we provide brief explanations for all fourteen non-null models, including references to full explanations of existing models. Unless otherwise noted, those seven models were implemented exactly as proposed in the original papers. Throughout, we use C to denote potential causes, and E to denote the effect. We will be analyzing data from experiments with only one potential cause, and so we will largely focus on that special case.

Existing Models

(1) Augmented Rescorla-Wagner model: A prime example of associationist learning theory, the Rescorla-Wagner model posits that individuals learn causal relationships via an error correction process (Rescorla & Wagner, 1972). For the present case of only one potential cause, the learned strength of C at time t is given by V_C^t. There is also assumed to be an always-present background cause with strength V_B^t. These strengths are used to predict the next case, and then are updated based on the discrepancy between one's prediction and the observation. More precisely, $V_X^{t+1} = V_X^t + \Delta V_X^t$ where:

$$\Delta V_X^t = \alpha_{X\delta(X)}\beta_{\delta(E)}\left(\lambda\delta(E) - \textstyle\sum_{\text{all causes } Y}\delta(Y)V_Y^t\right)$$

$\delta(X)$ is an indicator function for whether X is present; $\alpha_{C0/1}$, $\alpha_{B0/1}$, and $\beta_{0/1}$ are learning rate parameters for the absence or presence of C, B, and E, respectively; and λ denotes the salience of E. All parameters are assumed to be determined by features of the stimuli, but are free parameters in our model-fitting exercise.

(2) Probabilistic Contrast: The probabilistic contrast model argues that an individual's causal strength judgments are given by a standard measure of association (Cheng & Novick, 1992), and also describes the long-run equilibria of the (augmented) Rescorla-Wagner model (Danks, 2003). There are both unconditional and conditional probabilistic contrasts when there are multiple potential causes. For the present one-cause case, these quantities collapse together and we have:

$$\Delta P = P(E \mid C) - P(E \mid \neg C)$$

There are no free parameters for probabilistic contrast.

(3) Power PC: The power PC or causal power theory assumes that individuals behave as if relationships between causes and effects cannot be directly observed, but rather must be inferred from the observable data (Cheng, 1997). The original theory provided only long-run or asymptotic predictions. In the present analysis, we use a dynamical version of the power PC theory (i.e., the equilibria of the dynamical version are the same as the long-run power PC predictions) that provides case-by-case predictions (Danks,

Griffiths, & Tenenbaum, 2003). For single-cause situations, the simplified update equations are:

$$\text{If } V_X^t \geq 0\text{: } \Delta V_X^t = \alpha_{X\delta(X)}\beta_{\delta(E)}(\lambda\delta(E) - (V_B^t + \delta(C)V_C^t - V_B^t\delta(C)V_C^t))$$

$$\text{If } V_X^t < 0\text{: } \Delta V_t = \alpha_{X\delta(X)}\beta_{\delta(E)}(\lambda\delta(E) - V_B^t(1 - \delta(C)V_C^t))$$

(4) Proportion of Confirming Instances (pCI): White's (2003a, 2003b) pCI model posits that individuals track the evidence that logically confirms vs. disconfirms the hypothesis that C causes E. Cases in which C and E match (i.e., both occur or both fail to occur) provide confirmation of that hypothesis; cases in which they differ disconfirm that hypothesis. The pCI theory holds that causal judgments are given by the difference in total probability between these groups of evidence:

$$\text{pCI} = [P(C\&E) + P(\neg C\&\neg E)] - [P(C\&\neg E) + P(\neg C\&E)]$$

(5) Belief Adjustment Model: A generalization of the pCI theory holds that individuals only update their beliefs when asked to make a new judgment, and they do so by error-correcting (in terms of pCI) since their last judgment (Catena, Maldonado, & Candido, 1998). More precisely, suppose that $J_1, \ldots, J_n$ denote the causal judgments made by the learner, and let $pCI(k, k{+}1)$ denote the pCI for all cases between J_k and J_{k+1}. If γ is a learning rate parameter, then we update on each new judgment as:

$$J_{n+1} = J_n + \gamma(pCI(n, n{+}1) - J_n)$$

For the experiments analyzed in this paper, participants gave explicit judgments after every trial, so the update equation is well-defined for all individuals.

(6) Causal Support: The causal support model (Griffiths & Tenenbaum, 2005) postulates that causal inference involves weighing the evidence one has in favor of various possible causal graphs. Given only one potential cause (with an always-present background cause), there are two potential graphs: $\mathbf{G}_1 = \{C \rightarrow E \leftarrow B\}$; and $\mathbf{G}_2 = \{C \quad E \leftarrow B\}$. Causal support is then defined as:

$$\text{Causal support} = log(\, P(D \mid \mathbf{G}_1) \,/\, P(D \mid \mathbf{G}_2)\,)$$

To convert this quantity into the [-1, 1] interval, we transform causal support through a sigmoid function (with free parameters α, β to control the shape), and make preventive causes negative.

(7) Sequential Bayesian Theory: Standard Bayesian theories are order-invariant with regards to the evidence: they predict the same response regardless of the order in which the data are perceived. In order to better account for a variety of experimental order effects, a modified sequential Bayesian update procedure was developed and applied to similar sorts of data as we consider here (Lu, *et al*., 2016). Broadly, the model involves a two-step process: (1) at time t, the learner generates an expected distribution for the causal weights for the next trial; (2) given the observed D_{t+1}, a correction step uses Bayes rule to update the prediction distribution in light of the observed data. Mathematically, if w^t denotes the causal strengths/weights at time t and M denotes the causal generative model that the learner assumes, then we update as:

$$\text{Step (1): } P(w^{t+1} \mid D^t, M) = \int dw^t\, P(w^{t+1} \mid w^t)\, P(w^t \mid D^t, M)$$

$$\text{Step (2): } P(w^{t+1} \mid D^{t+1}, M) = P(D^{t+1} \mid w^{t+1}, M)P(w^{t+1} \mid D^t, M) \,/\, P(D^{t+1} \mid D^t, M)$$

Novel Models

The previous seven existing models represent a wide range of approaches to the cognitive task of causal inference. Nonetheless, we developed seven additional models by drawing on inspiration from analogous theories and ideas from other domains of cognitive science, particularly about the ways that resource limitations in attention and/or memory can yield heuristic learning methods that nonetheless perform relatively well.

(8) Bayesian Correlation Optimization: A straightforward causal inference strategy is simply to compute the correlation coefficient between C and E (using Bayes Theorem to provide regularization). Specifically, we treat C and E as binary (Bernoulli) random variables. After each case, we compute the correlation coefficient for the sequence of sums of C and E with highest posterior probability.

(9) Moving *k*-Window: The moving k-window model posits that individuals do not perform inference on the entire sequence of cases, but only on the most recent k cases. This type of input restriction could, in theory, be used with any of the other models, though it would not substantively change predictions of the dynamical models. We focused on a version that implements pCI over the past k cases for two reasons. First, the Belief Adjustment Model already incorporates similar ideas about memory bounds. Second, that theory is computationally quite simple which intuitively coheres with imposition of a memory constraint. More specifically, the causal judgment at t is given by:

$$V_C^t = \text{pCI}(t - k, t)$$

where $pCI(i, j)$ denotes pCI computed over cases i through j. For purposes of model fitting, we considered $k \in [1,10]$

(10) Win-Stay, Lose-Shift: The win-stay, lose-shift model postulates that individuals increase their judgment of the causal strength between C and E if the most recent case fits with their current belief about the direction of the causal relation (i.e., if they see a "win"). If the most recent case diverges from their current judgment (i.e., a "loss"), then they switch their judgment from generative to preventive (or vice versa). The basic idea of Win-Stay, Lose-Shift has found support in other experiments (Bonawitz, Denison, Gopnik, & Griffiths, 2014). We adapt the idea to causal strength learning using four conditional update steps:

- If $V_C^t \geq 0$ and $\delta(C) = \delta(E)$: $V_C^{t+1} = V_C^t + (1 - V_C^t)/2$
- If $V_C^t \geq 0$ and $\delta(C) \neq \delta(E)$: $V_C^{t+1} = -0.5$
- If $V_C^t < 0$ and $\delta(C) = \delta(E)$: $V_C^{t+1} = 0.5$
- If $V_C^t < 0$ and $\delta(C) \neq \delta(E)$: $V_C^{t+1} = V_C^t - (1 + V_C^t)/2$

(11) Rescorla-Wagner with exponential decay: One potential shortcoming of the Rescorla-Wagner and (dynamical) power PC models is that they only stabilize on a particular long-run value in special contexts. In most settings, they continue to vary around their equilibrium values without ever converging (Danks, 2003), in contrast with observed human behavior. We thus considered a generalized version of Rescorla-Wagner in which the learning rate exponentially decays towards zero to ensure convergence. Mathematically, we multiply the ΔV_X^t terms by $\mu = e^{-\lambda t}$, where λ is a free parameter that controls the speed at which the learner converges on a stable judgment. This model thus adds one free parameter to the unmodified Rescorla-Wagner model.

(12) Power PC with exponential decay: We similarly modified the dynamical power PC theory by multiplying the ΔV_X^t terms by $\mu = e^{-\lambda t}$ to ensure converge on a stable judgment as the number of cases increases.

(13) Rescorla-Wagner with stability of beliefs: The previous two models use a relatively blunt modification to ensure stabilization, as they assume that the learning rate converges to zero independently of the learner's beliefs. An alternative approach would be to base the learning rate on (the inverse of) the stability of the learner's recent judgments. That is, if the learner has (or has not) significantly changed her beliefs after recent evidence, then her learning rate should be relatively large (or small). Similar computational ideas are employed in many machine learning algorithms to determine when learning has largely stopped or otherwise adapt the learning rate(s). Let σ equal the average change in causal strength predictions over both C and B in the last k cases. We then multiply the ΔV_X^t terms by $\mu = \max(\gamma\sigma, 1/\sqrt{t})$, where γ is a rescaling parameter to ensure large σ values do not lead to abnormally large changes in strength judgment (and the second term in the max() function ensures that the learner does not converge too fast). Note that, in contrast with the previous two models, this model will not necessarily converge on a stable judgment if the learning environment is sufficiently non-stationary.

(14) Power PC with stability of belief: Similarly, we modified the dynamical power PC model by multiplying those ΔV_X^t terms by $\mu = \max(\gamma\sigma, 1/\sqrt{t})$.

Evaluating the Models

We fit data from two experiments, each with multiple conditions, that were first reported in Danks & Schwartz (2005, 2006). In both experiments, participants saw a sequence of binary cause and effect cases (using a cover story involving plants and skin rashes). After each case, participants were asked to estimate the strength of the cause-effect relationship on a [-100, +100] scale. Crucially, within each sequence of cases, the causal relationship (if any) changed to the opposite valence halfway through the sequence (without any notice to the participant). These sequences were non-stationary: all had the same overall statistics with $P(E \mid C) = P(E \mid \neg C) = P(C) = 0.5$, but most half-sequences had significant C-E correlations. They also present particularly challenging data for the various models, as participants' causal strength beliefs may exhibit significant variability over the course of a sequence. (Interestingly, no participants in either experiment reported conscious awareness of the changes in causal strength nor the non-stationarity, though a few reported that the sequences seemed a bit "odd.")

In Danks & Schwartz (2005), participants saw six sequences with 8 (twice), 16, 32, 48, and 80 cases. Each sequence had (i) strong positive correlation in the first half and strong negative in the second half; (ii) the opposite structure; or (iii) no correlation throughout the whole sequence. Sequence-type was counterbalanced within-participant but across lengths; for each participant, there are judgment curves for six different conditions. This dataset contains 51 participants. In Danks & Schwartz (2006), all participants saw the same five 40-case sequences, one with zero correlation and four with {weak correlation, strong correlation} × {positive correlation first, negative correlation first}. Hence, there are five different conditions for each participant. This dataset contains 40 participants. The original experiments found only limited order effects in terms of average final ratings; the mean strength ratings largely converged back to zero in each condition (though as shown in Figure 1, there was substantial variability along the way).

To fit specific parameterized models to participant data, we first generated model predictions for each possible condition by varying model parameters in a grid structure (i.e., all possible ways of varying parameter values across fixed ranges in fixed steps). Our models were relatively simple in number and range of parameters, so there was no need to use more sophisticated parameter estimation methods. We then determined, for each participant in each condition, the parameterized model with the lowest sum of squared error (SSE) for the actual participant judgments. We also determined the parameterized model that best fit (i.e., minimized SSE) each participant's judgments across all five or six conditions. This latter analysis enables us to assess whether participants seemed to use a stable learning strategy across multiple conditions of a single experiment.

One significant concern about our use of SSE is that it does not correct for the number of parameters in each model, in contrast with other model selection measures that penalize models with more free parameters (e.g., AIC, BIC). For many of our models, however, it is not clear exactly how to count the number of parameters. For example, the parameter values in the augmented Rescorla-Wagner model are typically thought to constrain one another in substantive ways (e.g., the learning rate for absent cues should be smaller than for present cues), and so we cannot simply use the number of named parameters in a straightforward way. In light of this complexity, we opted for the admittedly blunter measure of SSE, with the recognition that more complex models will likely fare better.

Individual Differences Analysis

The most basic level of analysis is simply which model fit the participant-level data best; see Figure 2 for that result, where we exclude models that fit less than 5% of the conditions. For both datasets, variants on the associationist models—either Rescorla-Wagner or dynamical power PC—dominated the model-fitting competition. Participant-conditions were largely best fit by some kind of error-correction procedure, coupled with some modification to help produce belief convergence.

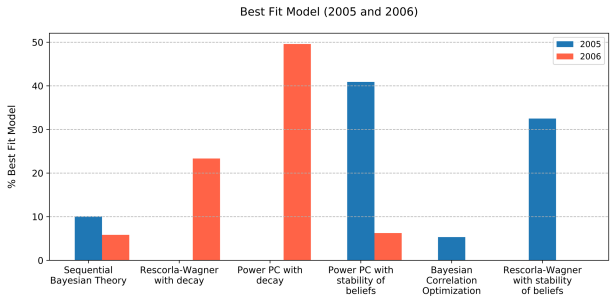

Figure 2: Distribution of best-fitting models

However, this apparent uniformity masks some significant within-participant diversity. A simple measure of the stability of a participant's strategy is the number of different model-types that were the best fits across the five (for 2005 data) or six (for 2006 data) conditions. For this analysis, we ignore the actual parameter values and focus solely on the model-type. Table 1 gives the percentages of participants who had the given number of different model-types across their conditions.

Table 1: Distribution of distinct model-types.

# of model-types	% of 2005 participants	% of 2006 participants
1	3.9	10.0
2	21.6	25.0
3	52.9	57.5
4	15.7	7.5
5	5.9	0.0
6	0.0	

The within-participant diversity is limited: *no* participants (in either experiment) had different best-fitting model-types for every condition in that experiment. Every participant had at least two conditions for which the same model-type provided the best fit. At the same time, relatively few participants (3.9% in the 2005 data; 10.0% in the 2006 data) had the same best-fitting model-type for *every* condition in their experiment. Based on the by-condition analysis, most participants appeared to use a mix of model-types in their causal learning.

A natural question is whether there is any pattern to the distributions of model-types. Figures 3 and 4 show the model-type distributions by trial length (2005 data) and by sequence type (2006 data). These Figures reveal that the Sequential Bayesian Theory performs better on shorter sequences (2005), and on unbiased sequences (2006). The latter finding is unsurprising, given that the order-invariance of "normal" Bayesian models. Moreover, Power PC-based methods do better for more "extremal" conditions, either longer sequences or stronger causes.

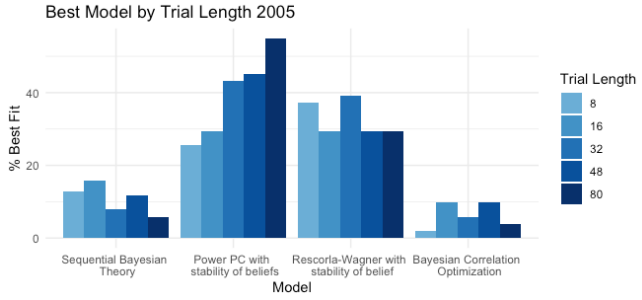

Figure 3: Best-fitting models by trial length (2005)

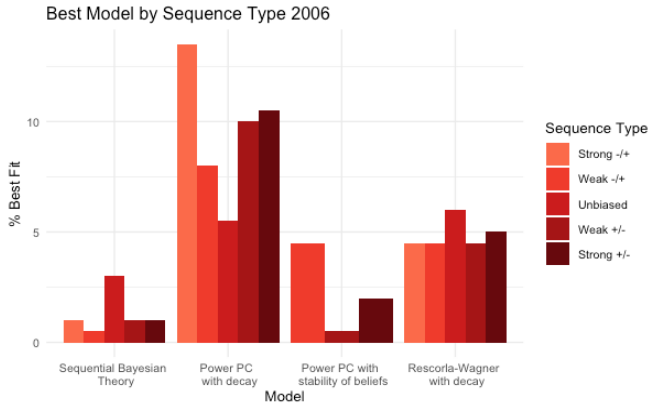

Figure 4: Best-fitting models by sequence type (2006)

Of course, "best-fitting" does not mean "good-fitting"; the best model of a participant's judgments might actually be a very poor model. We say that a model is "reasonable" if the average error per judgment/datapoint is less than five (since judgments were on a [-100, +100] scale). This measure (rather than, say, total error) allows comparison of model performance across sequences of different lengths. Qualitatively similar results were obtained for different thresholds; in particular, allowing slightly larger average errors to count as "reasonable" did not substantively improve matters. Having said that, one concern about this criterion for 'reasonable' is that it focuses on pointwise errors, rather than overall patterns. A model that predicts the exact learning trajectory offset by 10 would, on this criterion, count as unreasonable even though it perfectly captured the pattern.

Table 2 shows the percentage of participant-conditions for which there was a reasonable model, separated by condition. There is a clear impact of condition: reasonable models were more likely to be found for shorter sequences and those with

weaker (or zero) half-sequence correlations. More importantly, the overall low percentage of reasonable models suggests that many participants were learning differently than proposed by *any* of the models that we included.

Table 2: Frequency of "reasonable" models.

2005		2006	
condition	% reasonable models	condition	% reasonable models
8	46.1	Strong -/+	20.0
16	37.3	Strong +/-	32.5
32	39.2	Unbiased	42.5
48	31.3	Weak -/+	40.0
80	19.6	Weak +/-	55.0

We can also focus on the single model that provided the best fit for each participant across all of their conditions. Figure 5 shows the best-fitting single model-type for each participant; again, modified associationist models perform the best, with power PC variants leading the way.

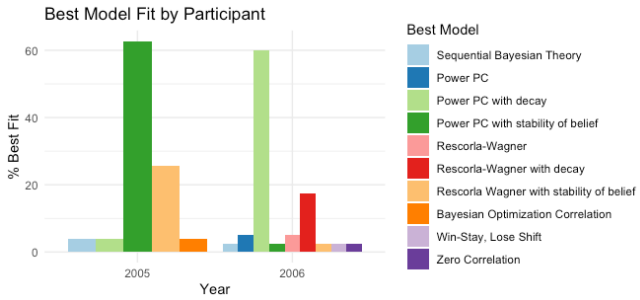

Figure 5: Distribution of best-fitting single models

The best-fitting single models can potentially help us to understand the amount of within-participant variation. In particular, the total SSE for the best-fitting single model must be greater-than-or-equal-to the sum of SSEs from each of the best-fitting per-condition models. We can thus divide each single-model SSE by the sum of per-condition model SSEs to gain an understanding of how much worse the single model performs (where a higher number suggests that this participant is more likely to be using different strategies across conditions). For the 2005 participants, the median ratio was 2.24; for the 2006 participants, the median ratio was 3.78. That is, the single-model SSE was 2-3 times worse than the sum of the per-condition model SSEs. These results thus provide additional support for the earlier conclusion that most participants seem to be using more than one strategy across the various experimental conditions.

Next Steps & Conclusions

This analysis is a first step towards shedding new light on individual causal learning strategies. We found non-trivial diversity both within- and between-participants; few participants seem to be using the same learning strategy in all conditions, and no model-type was found to be universal. Nonetheless, certain model-types were disproportionately represented at both levels of analysis. In particular, error-correction models (modified to increase stability of long-run judgments) consistently were the best-fitting models, whether on a per-condition or per-participant basis. These models were also the most complex, though, so future work should explore performance measures that penalize model complexity (though *pace* the earlier observations about the difficulty of counting parameters for some models).

At the same time, the overall model fits were surprisingly poor. In almost every condition, fewer than half of the participants were best-fit by a model with an average per-datapoint error less than five. These weak model fits are all the more surprising given the strong predictive performance previously reported for many of the existing models. Predictions of several models have been shown to be highly correlated ($\rho > 0.9$) with average participant judgments, and so we expected that they would perform well in our analyses.

One possible explanation is that these models have largely been compared to people's long-run, stable judgments after observing many cases. In contrast, our analysis tried to fit these models to case-by-case participant judgments in response to observations from non-stationary distributions. The experimental task may have prompted participants to use different types of learning strategies than are used in other types of experiments. Alternately, theories intended for long-run predictions might not translate well to case-by-case judgments (though only models (2), (4), and (6) seem potentially restricted to long-run predictions).

Another possible explanation is that theories of average causal learning behavior are simply not good theories of any particular individual's causal learning behavior. As has been shown in other settings, the best model of average performance can be quite different from every individual's personal best model (e.g., Brown & Heathcote, 2003; Myung, Kim, & Pitt, 2000). Causal learning may provide another such example. For either potential explanation, however, we must conduct significant further inquiry into the ways that individual people learn causal strengths in complex sequential environments.

References

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology, 74*, 35–65

Brown, S., & Heathcote, A. (2003). Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt. *Memory & Cognition, 31*, 656-661.

Catena, A., Maldonado, A., & Candido, A. (1998). The effect of the frequency of judgment and the type of trials on covariation learning. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 481–495.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*, 109–121.

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67–74). Cambridge, MA: The MIT Press.

Danks, D., & Schwartz, S. (2005). Causal learning from biased sequences. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual meeting of the cognitive science society* (pp. 542-547). Mahwah, NJ: Lawrence Erlbaum Associates.

Danks, D., & Schwartz, S. (2006). Effects of causal strength on learning from biased sequences. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 1180-1185). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive science*, *31*(5), 765–814.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135–163.

Lee, J. C., & Lovibond, P. F. (2020). Individual differences in causal structures inferred during feature negative learning. Quarterly Journal of Experimental Psychology. https://doi.org/10.1177/1747021820959286

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science, 40*(2), 404-439.

Michotte, A. (1946). La perception de la causalité. Louvain: Institut Superior de Philosophie, 1946. English translation of updated edition by T. Miles & E. Miles, *The Perception of Causality*, Basic Books, 1963.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power-law artifact: Insights from response surface analysis. *Memory & Cognition, 28*, 832-840.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin and Review, 14*, 577–596.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*. doi:10.1037/a0031903

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

White, P. A. (2003a). Making causal judgments from the proportion of confirming instances: The pCI Rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*(4), 710–727.

White, P. A. (2003b). Causal judgement as evaluation of evidence: The use of confirmatory and disconfirmatory information. *The Quarterly Journal of Experimental Psychology*, *56A*(3), 491–513.