Inducing preference reversals by manipulating revealed preferences

Harish Balakrishnan (harishb@iitk.ac.in)
Cognitive Science, IIT Kanpur, India

Shobhit Jagga (shobhitj@iitk.ac.in)
Computer Science, IIT Kanpur, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Computer Science, IIT Kanpur, India

Abstract

It is currently difficult to test the validity of existing explanations for the emergence of context-dependent preference reversals. This is because these explanations are generally placed at the level of the process of evidence accumulation, and across experimental paradigms, this process is unobservable. In this paper, we propose a new experimental paradigm for eliciting preference reversals, wherein the process of evidence accumulation is significantly observable. Over a series of experiments, we successfully induce preference reversals for arbitrary stimuli by showing participants sequences of stimuli comparisons with pre-determined outcomes. Our findings partially support the view that context-sensitive assimilation of a history of ordinal comparisons is sufficient to explain classic context effects.

Keywords: preference reversals; decisions from experience; preference formation

Introduction

Preference reversals occupy a particularly interesting niche in research at the interface between psychology and economics (Rabin, 1998). The earliest conceptualization of a preference reversal can be traced to Luce & Raiffa's fastidious diner, who initially prefers salmon to steak off of a restaurant menu, but changes his mind and orders steak instead when the waiter tells him that the day's special is frog legs (Luce & Raiffa, 1957). While whimsical, this story points directly to a core objection to the tenability of option-specific representations of value in the mind (Srivastava & Schrater, 2015). If the diner prefers salmon to steak to begin with, why does the introduction of an additional item shift their preference to steak?

In Luce & Raiffa's original explanation, this happens because the observation that frog legs were on the menu raised the diner's expectation of the quality of the restaurant, causing him to change his mind and ask for steak, which is harder to cook right than salmon (Luce & Raiffa, 1957). In other words, the diner infers something about the generative process underlying the options from the set of options, and then uses his understanding of the generative process to construct his preference. Preference reversals and more generally, all such context effects, are deeply interesting because they uncover the existence of such sophisticated inferences underpinning the simple act of choosing between items (Srivastava & Schrater, 2015).

There have been a number of successful attempts to reproduce all three of these classic preference reversals within unitary computational models (Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2004; Bhatia, 2013; Shenoy & Yu, 2013; Srivastava & Schrater, 2015), all with different possible interpretations of the potential causes for each of the effects, and different caveats for their appearance. Given this proliferation of widely divergent possible explanations, it becomes important to differentiate them based on criteria beyond qualitative reproduction of the effects.

Interestingly, in recent years, these context effects have been documented in paradigms beyond affective preferences, such as inductive reasoning (Trueblood, 2012) and perceptual judgement (Trueblood, Brown, Heathcote, & Busemeyer, 2013). Such demonstrations call into question unitary accounts of these effects that place their explanations on economic assumptions, such as the loss aversion assumption in the LCA model (Usher & McClelland, 2004) and the market value discovery assumption in Shenoy & Yu's Bayesian observer model (Shenoy & Yu, 2013).

Further, context effects have been documented for multiple non-human animals: preference reversals induced by change of frame for capuchin monkeys (Lakshminarayanan, Chen, & Santos, 2011), context-dependent foraging decisions in hummingbirds (Bateson, Healy, & Hurly, 2003), and perhaps most impressive, the elicitation of an asymmetric dominance effect in food location preference observed in the acellular protist *physarum polycephalum* (Latty & Beekman, 2011). These observations, the last one in particular, suggest that the true explanations for these context effects likely lie in simple information processing mechanisms, such as the ones proposed in decision field theory (Roe et al., 2001), evidence accumulation based on ordinal comparisons (Ronayne & Brown, 2017; Noguchi & Stewart, 2018) or through inference based on ordinal comparisons (Srivastava & Schrater, 2015).

But while simple information accumulation-based explanations are promising, they have not been directly tested. This is, in large part, because information accumulation explanations make claims about the process by which evidence is accumulated, and the process of valuation, be the paradigm affective (Huber, Payne, & Puto, 1982), inferential (Trueblood, 2012) or perceptual (Trueblood et al., 2013), is hidden from the experimenter's view.

For instance, decision field theory explains the asymmetric dominance (attraction) effect as a result of a negative preference created for the inferior decoy introduced, which prop-

agates through a negative inhibitory link to the dominating option, causing it to increase in valuation (Roe et al., 2001). None of these postulated intermediate calculations are observable, making it impossible to render any judgement on the validity of the proposed mechanism as the correct explanation of the effect. To properly evaluate whether evidence accumulation accounts are adequate to explain preference reversals, a new experimental paradigm is needed, wherein the valuation process is within the experimenter's view.

In this paper, we present such a paradigm, and attempt to elicit the three classic preference reversals (attraction, similarity and compromise) in it, based on predictions made by the ordinal comparison-based account of preference reversals (Srivastava & Schrater, 2015).

Methods

Manipulating revealed preferences

In this paradigm (illustrated in Figure 1), participants are shown pairwise comparison presentations of different options with the preferred option revealed in all cases, followed intermittently by preference input solicitations. A *comparison presentation* is a trial in which the relative superiority of an option is demonstrated as an ordinal comparison. Each comparison presentation is an animation that lasts for two seconds. After a fixed number of such trials, participants are asked for their preferences between the two options. These preference inputs constitute their baseline preferences for the two original options. Next, we present similar comparison presentation trials pairing the target option with the decoy, and at the end of the presentation sequence, elicit baseline preferences for all relevant pairs of options. In Phase 2, participants are asked for their final preferences with a short break ($\sim$ 90 to 180 seconds) in between phases 1 and 2. The order of presentation of comparison sequences is random across participants within conditions.

Inducing preference reversals

Prior literature agrees that different placements of the decoy (see Figure 2) relative to the target and the competitor options results in different types of preference reversals. In the attraction effect, the decoy is an asymmetrically dominated option that makes the target option appear more attractive than before. In the compromise effect, the decoy is an extreme option that increases the desirability of the target option by making it appear as a compromise between the competitor and itself. In the similarity effect, the decoy is very similar to the competitor option, and it makes the target option appear more salient than before, thereby increasing its final preference share.

This paradigm, thus, presents value signals in sequence much as in value psychophysics (Tsetsos, Chater, & Usher, 2012) and the experimental paradigm used by Ronayne and Brown (2017). The novelty of our paradigm is that we use races observed over time in an animation to present evidence for attribute-level superiority or inferiority, rather than leveraging numeric labels to present the same information. Thus for example, in the sample experiment illustrated in Figure 1, horses' speeds are not presented as numeric quantities, but observers can see over multiple trials that one of the horses tends to win more races than the other. Where the items have two attributes, we show sequences of races leveraging both attributes separately, as we detail further below. In all the cases described below, the specific sets of pairwise comparisons used are obtained from the ordinal comparison model's suggestions, as being likely to induce preference reversals (Srivastava & Schrater, 2015).

The principal value of shifting from presenting numeric attribute values to explicit ordinal comparisons is that whatever evidence for the superiority or inferiority of an item exists on any attribute dimension is observable to both the observer and the experimenter throughout the experiment, unlike in value psychophysics, where some underlying utility function must be assumed (Tsetsos et al., 2012). Thus, if preference reversals are obtained within this paradigm, the experimenter will clearly know the minimal amount of information observers need to construct valuations in ways that yield preference reversals.

Attraction Effect

The attraction (asymmetric dominance) effect appears to be the easiest one to induce in non-human organisms, including notably slime mold (Latty & Beekman, 2011), contraindicating explanations that rely on associativity (Bhatia, 2013) and lateral inhibition (Roe et al., 2001). A simple explanation for this effect comes from the ordinal comparison accumulation account of (Srivastava & Schrater, 2015) who claim that the dominating option gains valuation by virtue of winning more comparisons via simple vote-counting. This theory further predicts that such an asymmetric dominance effect should hold independent of the number of attributes in the item set. To test this explanation, in addition to the conventional attraction effect setup using stimuli varying along two attributes, we also tested a version with unidimensional stimuli.

In both experiments, A was the target, B was the competitor and C was the decoy (see Figure 2). Phase 1 was divided into two blocks of comparison presentation trials, and baseline preferences were obtained after every five such trials. The sets of 5 pairwise presentations within each block and the trials within each set were randomized. The order of presentation of the choice sets in phase 2 was also randomized. Comparison sequences were designed to present the original options as approximately matched and the inferior decoy heavily dominated by the target option in terms of number of wins.

Experiment 1a - Stimuli with one attribute dimension: Participants were shown simulations of football matches between four teams - A, B, C, and D. Block 1 of phase 1 consisted of a set of ten matches of A versus B and a set of ten matches of C versus D in which all the teams win an equal number of matches and the net goal difference is zero. Block 2 consisted of a set of ten matches of A versus C, and

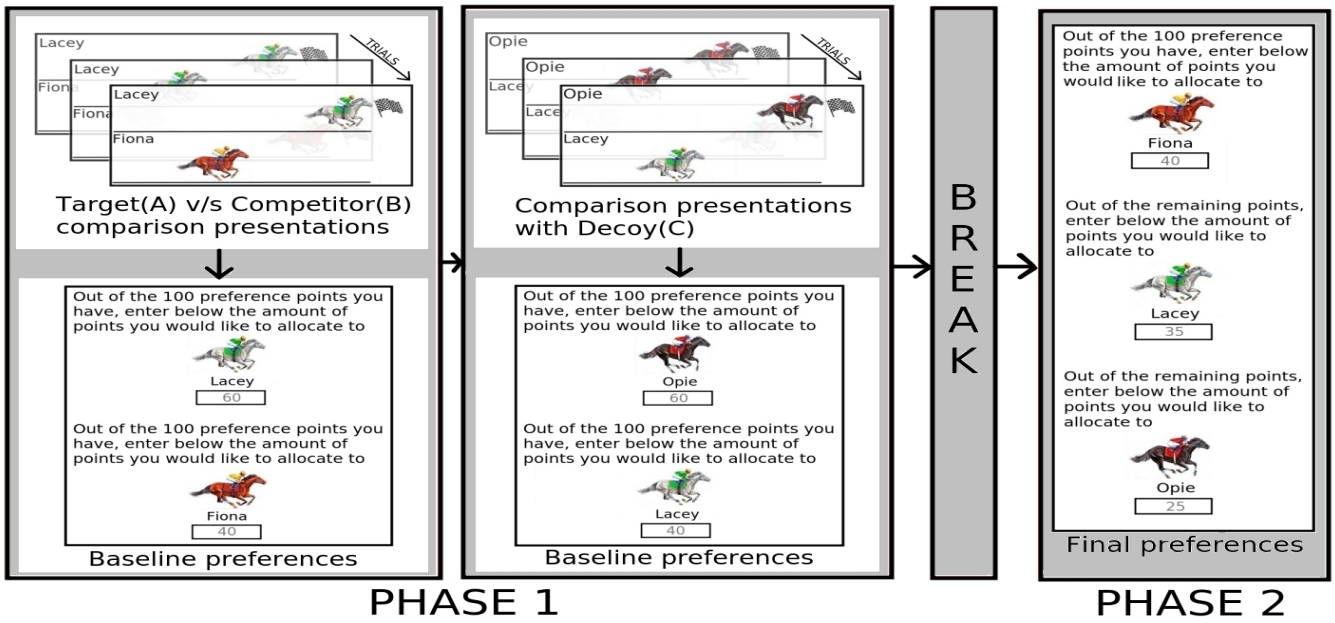

Figure 1: Sample experiment protocol. In phase 1, subjects were shown pairwise comparisons of the three options (horses *A*, *B*, and *C*) along two attribute dimensions (*race wins* and *money saved at maintenance*), without resorting to numerical representations for these attributes. A trial in phase 1 is an animation that compared two horses along one of the dimensions. After ten such trials, preference inputs were solicited from the participants. These preference inputs constitute their baseline preferences. A short break separates phases 1 and 2. In phase 2, participants were asked for their final preferences for all the three horses.

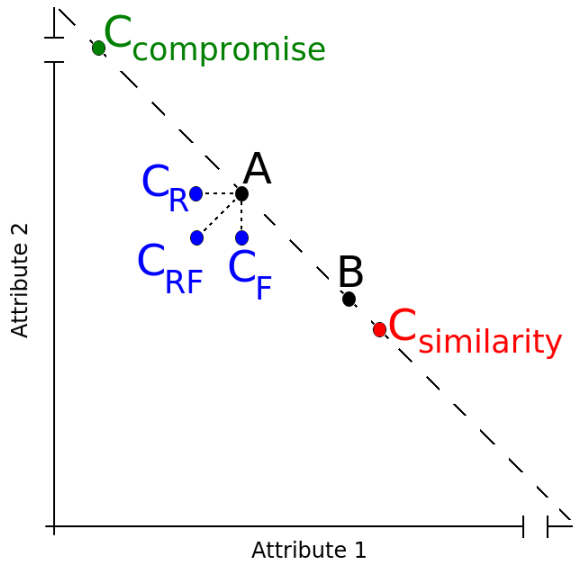

Figure 2: Different placements of decoy *C*, when *A* and *B* are the target and competitor options respectively. In all cases, the introduction of the decoy increases the preference share of *A* compared to *B* (Trueblood et al., 2013).

ten matches of *B* versus *D*. *A* always wins against *C* in the matches between *A* and *C*; *B* and *D* win an equal number of times in the *B* versus *D* matches. In phase 2, participants were sequentially asked for their final preferences for four pairwise choice sets - $\{A \text{ or } B\}$, $\{C \text{ or } D\}$, $\{A \text{ or } C\}$ and $\{B \text{ or } D\}$. Participants were asked to choose the better team in all choice sets.

Experiment 1b - Stimuli with two attribute dimensions Participants were shown pair-wise horse races between four horses - *A*, *B*, *C*, and *D*. Their performance varied along two attribute dimensions - race wins and maintenance costs after every race, represented as stacks of food. Both these attributes were seen simultaneously by the participants in every trial. The cover story implied the food costs were substantial. Block 1 of phase 1 consisted of a set of ten races of *A* versus *B* and a set of ten races of *C* versus *D* in which all the horses win five races each. In these races, the food stacks were either four or six in count and were the same for both the horses involved to emulate the initial conditions for the attraction effect ($A \sim B$). Block 2 consisted of a set of ten races of *A* versus *C* in which A wins eight times out of ten. Since a frequency decoy was used, the number of food stacks in the A versus C comparisons was also kept the same in any trial. In phase 2, participants were randomly assigned to two option choice contexts or three option choice contexts. In both the contexts, participants were shown all possible sets of choices and their preferences were obtained by asking them to split 100 prefer-

ence points between all probed options.

Compromise Effect

In this experiment, the target and the competitor options were determined dynamically after the first set of comparison presentations between *A* and *B*. The decoy placement was determined based on whether participants preferred *A* more than *B*, or *B* more than *A* initially. Phase 1 was divided into two blocks, and baseline preferences were obtained after every 10 comparison presentations. The sets within each block and the trials within each set were randomized. The order of presentation of the choice sets in phase 2 was also randomized. Here, comparisons sequences were designed to show the original options slightly dominating each other along different attributes, and the introduced decoy strongly dominating the competitor and weakly dominating the target, in terms of wins.

Participants were shown animations corresponding to pairwise benchmarking tests between three computer configurations that varied along two attribute dimensions - CPU performance and GPU performance. The relative superiority of options along these attribute dimensions was shown separately - with CPU-heavy computers shown to perform a compute-intensive test sooner, and GPU-heavy computers shown to render complex graphics quicker. Block 1 of phase 1 consisted of ten trials of *A* versus *B* in each of the two attribute dimensions. *A* beats *B* six out of ten times in the CPU performance tasks, and *B* beats *A* six out of ten times in the GPU performance tasks. The target and competitor options were determined after block 1. If the target option was *A*, then GPU performance was the dominant dimension, and if the target option was *B*, then CPU performance was the dominant dimension. Block 2 of phase 1 consisted of ten trials of target versus decoy comparisons and ten trials of competitor versus decoy comparisons in each of the two attribute dimensions. In the target-decoy comparisons, the target wins seven times in the dominant dimension, and the decoy wins six times in the other dimension. In the competitor-decoy comparisons, the competitor wins eight times in the dominant dimension, and the decoy wins eight times in the other dimension. Preference inputs were obtained in the form of preference points as above.

Similarity Effect

Experiment 3 - Stimuli with two attribute dimensions: The design of experiment 3 was similar to experiment 2, except for the following changes. Participants were shown horse races between three horses - *A*, *B*, and *C*. Their performance varied along two attribute dimensions - race wins and money saved in maintenance costs, which the cover story implied were substantial. The trials from two sets interleaved dimensional prominence i.e., a set in experiment 3 consisted of five pairs of trials, and a pair consisted of a race comparison trial followed by a savings trial. In the target-decoy comparisons, the decoy wins seven times in the dominant dimension, and the target wins seven times in the other dimension. In the competitor-decoy comparisons, both the options were shown to be superior five times each in both the dimensions. Preference inputs were obtained in the form of preference points as above.

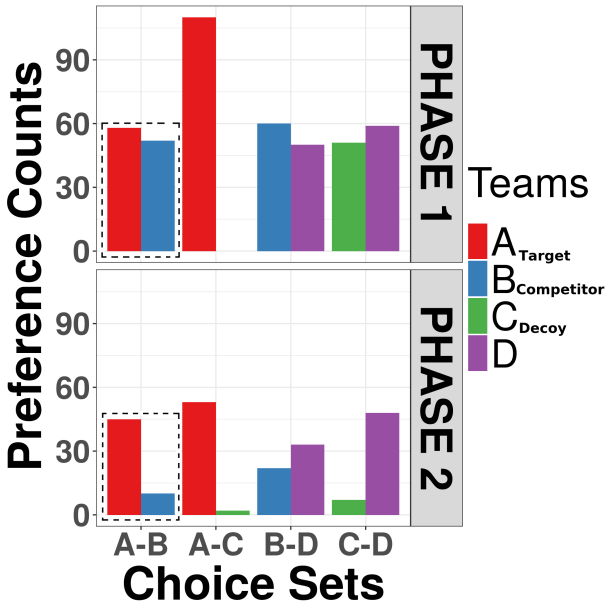

Figure 3: Results of experiment 1a. The first plot (from top) shows the participants' initial preferences in phase 1. The second plot shows their final preferences.

Sample

University students volunteered to participate in our experiments and were paid for participation. All study procedures and methods were reviewed and approved by an IRB.

Results

At the cohort level, we report results in terms of cumulative preference share in favor of all tested options over all participants. These results are depicted graphically via Figures 3-5.

Additionally, as a within-subject analysis, we performed McNemar's test over the counts of the number of participants choosing the target over the competitor before the introduction of the decoy and after the introduction of the decoy. Table 1 summarizes the results of this analysis.

Attraction Effect

In the attraction effect, initially, both the target and the competitor options are preferred almost equally. After the introduction of the asymmetrically dominated decoy, the final preference for the target is increased considerably.

Experiment 1a Fifty-five university students participated in this experiment. Figure 3 shows the cumulative preference responses of the entire cohort. Initial preference responses indicate almost an equal preference for both the target and the competitor. After the introduction of the decoy, the target's

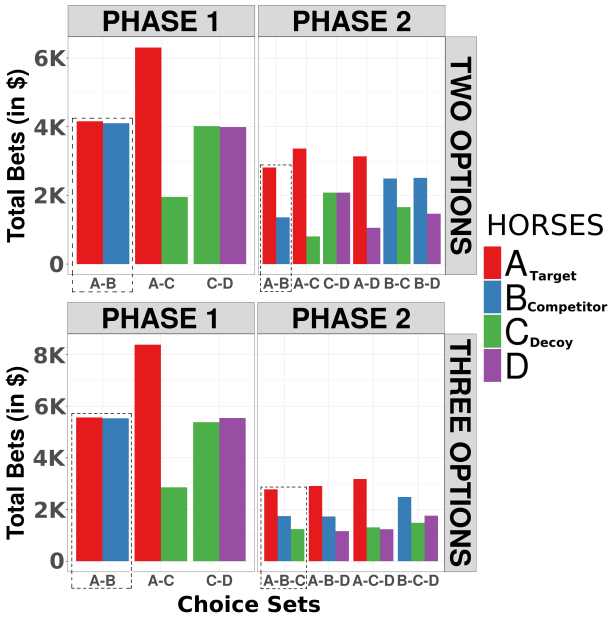

Figure 4: Results of experiment 1b. The plots in the first row (from top) correspond to the two-option final choice context condition, and the plots in the second row correspond to the three-option final choice condition.

preference share increased significantly, indicating a strong attraction effect.

Experiment 1b One hundred university students participated in this experiment. Forty-two students were randomly assigned to a binary final choice set condition, and the rest were assigned to a ternary final choice set condition. Figure 4 shows the cumulative preference responses of all participants. While the preference share of the target option does increase after the observation sequence, the shift in preference is not significant, as assessed by the McNemar test (see Table 1).

Compromise Effect

In the compromise effect, the competitor option is slightly preferred over the target option initially. After the introduction of an extreme decoy, the preferences are shifted towards the target option finally.

Thirty-four university students participated in the experiment. Figure 5 sub-figure A shows the cumulative preference responses of all the participants. Since the compromise effect requires the placement of the decoy such that the previously dominated competitor becomes a compromise option, there were two possibilities of decoy placement that were decided programmatically based on Phase 1 baseline preferences. Across both conditions, a significant compromise effect is observed (see table 1).

Similarity Effect

In the similarity effect, the target option is less preferred initially over the competitor option. After the introduction of the decoy that is similar to the competitor option, the preferences are shifted towards the target option finally.

Table 1: Results of the McNemar's test for all the experiments. Columns 2 and 3 show the number of subjects who chose Target(T) over Competitor(C) before and after the introduction of the decoy respectively.

Experiment	Discordant Cells		p-value
	Before: T>C	After: T>C	
1a	11	45	5.38×10^{-6}
1b (Two + Three)	43 (18+25)	62 (26+36)	0.07849
2 (CPU + GPU)	0	16 (4+12)	3.05×10^{-5}
3 (A bias + B bias)	0	18 (11+7)	7.63×10^{-6}

Fifty university students participated in the experiment. Figure 5 sub-figure B shows the cumulative preference responses of all the participants for the two cases. In both cases, the decoy option *C* makes the competitor option less salient and eats away a portion of its preference share, resulting in a comparatively lower final preference, indicating a similarity effect.

Model evaluation

In our paradigm, options are presented without using numerical attribute representations, and revealed preferences are manipulated through different sequences of comparison presentations over multiple trials. Since most existing models of preference reversals don't have a learning component, they cannot be applied to our data. However, the preference inference model mentioned above (Srivastava & Schrater, 2015) matches our paradigm well, and it can be applied to our data with a few parametric assumptions. Table 2 summarizes the model's trial-by-trial predictions using the same stimuli sequences we presented to human observers, assuming equal prior beliefs on different choice contexts involved. The model predictions match our data reasonably well for the attraction effect and the compromise effect, but not for the similarity effect.

Discussion

We present an experimental paradigm where observers learn to prefer arbitrary stimuli over others based on a sequence of ordinal comparisons. We find that introducing specific sequences of ordinal comparisons predicted by an observer model of preference inference (Srivastava & Schrater, 2015) to produce three classic preference reversals does do so to a considerable degree, but with interesting deviations. For instance, while an asymmetric dominance effect is clearly seen using unidimensional stimuli, it is not as clear using traditional two-dimensional ones. The compromise effect shows an unexpected dependence on attribute, suggesting potential

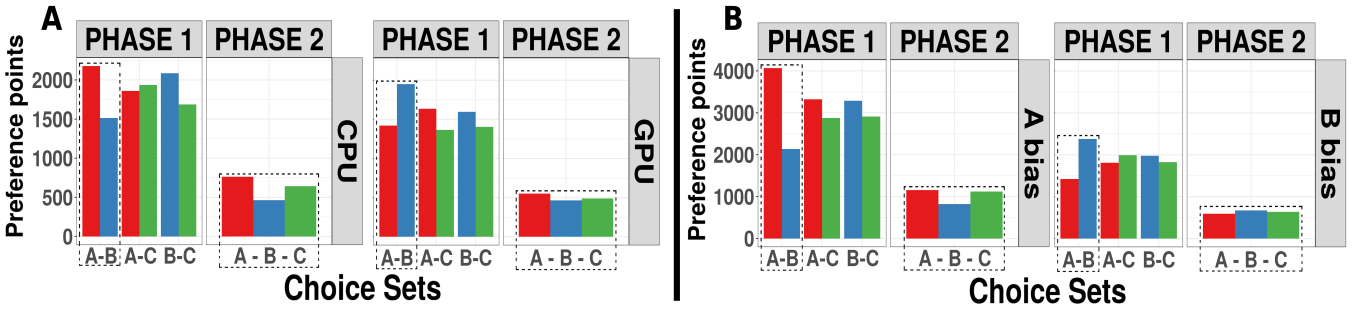

Figure 5: **(A)** - Results of experiment 2. The first two plots (from left) correspond to the case when the dominant dimension was CPU performance. The last two plots correspond to the case when the dominant dimension was GPU performance; **(B)** - Results of experiment 3. The first two plots (from left) correspond to the case when *horse A* was slightly preferred over *horse B*. The last two plots correspond to the case when *horse B* was slightly preferred over *horse A*; In both **A** and **B**, for each of the cases, the baseline preferences are shown in the first plot and the final preferences are shown in the second plot.

biases in participants' priors about the computer configuration cover story. Further investigation of the correspondence and deviations of behavior seen in this paradigm from this model presents a clear direction for future research.

While our results are consistent with the ordinal comparison account of preference reversals, in the absence of comparison with alternative accounts, it is premature to claim that they clearly differentiate it from alternatives. In particular, the ordinal comparison observer model, like several Bayesian observer models, does not in itself predict reversals, but rather serves as a container for assumptions about environmental influences, that ultimately provide the explanation for observed effects (Srivastava & Schrater, 2015). The ordinal comparison model's explanation for the similarity effect, for example, is identical to the one seen in decision field theory - that the decoy steals wins from the competitor, but not from the target (Roe et al., 2001), merely implemented differently. Thus, our current results cannot be used to support one model's case over another. At most, they can be used to argue in favor of simpler sequential information accumulation accounts of preference reversals, such as the ones discussed in Roe et al. (2001), Srivastava and Schrater (2015), Ronayne and Brown (2017) and Noguchi and Stewart (2018) over more complex theories that assume valuation-based mechanisms to produce such context effects.

This paper's main contribution is the demonstration that simple series of ordinal comparisons are sufficient to establish classic preference reversals, establishing that all that is really needed to see such effects is the ability to accumulate extremely coarse (even binary) task signals. This finding is congruent with the large variety of task domains in which preference reversals have been documented (Trueblood et al., 2013) and offers strong constraints on the nature of the mental representations that might apply across these domains to explain the emergence of such effects.

Table 2: Model predictions versus empirical observations along with the Matthews correlation coefficient(MCC) for all the preference reversal experiments.

		Empirical Observation	
		Preference reversal	Not a pref. reversal
Model Prediction	Preference Reversal	23,19,6,10	5,13,3,17
	Not a pref. reversal	11,14,10,8	16,54,15,15

MCC: Exp 1a=0.426, Exp 1b=0.385, Exp 2=0.236, Exp 3=0.023

Acknowledgments

Nisheeth Srivastava acknowledges funding from a DST CSRI research grant DST/CSRI/2017/334 and the Research I foundation.

References

Bateson, M., Healy, S. D., & Hurly, T. A. (2003). Context–dependent foraging decisions in rufous hummingbirds. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1521), 1271–1276.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, *120*(3), 522–543.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*, 90–98.

Lakshminarayanan, V. R., Chen, M. K., & Santos, L. R. (2011). The evolution of decision-making under risk: framing effects in monkey risk preferences. *Journal of Experimental Social Psychology*, *47*(3), 689–693.

Latty, T., & Beekman, M. (2011). Irrational decision-making in an amoeboid organism: transitivity and context-dependent preferences. *Proceedings of the Royal Society*.

Luce, D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. Wiley, New York.

Noguchi, T., & Stewart, N. (2018, July). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological review*, *125*(4), 512—544.

Rabin, M. (1998). Psychology and economics. *Journal of economic literature*, *36*(1), 11–46.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory:a dynamic connectionist model of decision making. *Psychological Review*, *108*(2), 370–392.

Ronayne, D., & Brown, G. D. (2017). Multi-attribute decision by sampling: An account of the attraction, compromise and similarity effects. *Journal of Mathematical Psychology*, *81*, 11 - 27.

Shenoy, P., & Yu, A. (2013). Rational preference shifts in multi-attribute choice: What is fair? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Srivastava, N., & Schrater, P. (2015). Learning what to want: Context-sensitive preference learning. *PLOS ONE*, *10*(10).

Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic bulletin & review*, *19*(5), 962–968.

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Association for Psychological Science*, *24*(6), 901–908.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. In *Proceedings of the national academy of sciences* (pp. 9659–9664). PNAS.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(3), 757–769.