Unrealized promise of joint modeling of choice and reaction time in improving representation learning

Russell Richie (drrichie@sas.upenn.edu)
University of Pennsylvania
3710 Hamilton Walk, Philadelphia, PA 19104 USA

Nehal Ajmal (najmal@vassar.edu)
Vassar College
124 Raymond Ave, Poughkeepsie, NY 12604 USA

Martin Hebart (hebart@cbs.mpg.de)
Max Planck Institute for Human Cognitive and Brain Sciences
Stephanstraße 1A, D-04103 Leipzig, Germany

Abstract

As mental representations are standardly thought to underlie all cognitive processes, a major goal of cognitive science has been to uncover representations. Methods for representation learning from behavioral data often model choice or reaction time data alone, but not jointly, leaving out potentially useful information. Here we develop two models of choice and RT in the odd-one-out task, including one based on the Linear Ballistic Accumulator. Parameter recovery simulations show joint modeling of choice and RT with LBA recovers representations more accurately than modeling choice alone with softmax. However, on two empirical datasets of images and words, joint models performed no better than choice-only models, despite a significant correlation of reaction time with two measures of similarity and choice difficulty in both datasets. We speculate on reasons for the unrealized promise of joint modeling of RT and choice in representation learning.

Keywords: similarity; computational modeling; reaction time; representation learning; concepts

Introduction

Mainstream cognitive science, in the Representational Theory of Mind tradition (Pitt, 2022), holds that mental representations, whether distributed or local/symbolic, underlie arguably all cognitive processes. Accordingly, cognitive scientists have invested significant effort developing methods for uncovering these representations, especially for distributed representations, which assume that objects and concepts are represented by – often high-dimensional – numerical vectors. Such methods include multidimensional scaling from Likert scale ratings and confusion matrices (Steyvers, 2002), sorting items into piles (Shepard & Cooper, 1992), spatially arranging items on a two-dimensional plane (Goldstone, 1994; Kriegeskorte & Mur, 2012; Hout, Goldinger, & Ferguson, 2013; Richie, White, Bhatia, & Hout, 2020), and, more recently, natural language processing (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and computer vision (Peterson, Abbott, & Griffiths, 2018). These methods have uncovered interpretable representations (Hebart, Zheng, Pereira, & Baker, 2020) that quantitatively predict behavior in generalization (Shepard, 1987), categorization (Nosofsky, 1984), semantic judgment (Richie, Zou, & Bhatia, 2019), similarity judgment (Richie & Bhatia, 2021), stereotyping (Bhatia, 2017), and more (Bhatia, Richie, & Zou, 2019).

A challenge in using these methods, however, is that they require significant dataset sizes for drawing conclusions that generalize beyond small, selective stimulus samples. For example, the number of trials needed to scale n items grows quadratically with pairwise ratings or confusion data, and cubically with odd-one-out judgments. The spatial arrangement method tries to circumvent this by scaling all items at once on a 2-d plane, but with item sets that are multi-dimensional, this requires repeated trials to allow respondents to focus on different item dimensions on each trial. The data intensity of all these methods makes it especially challenging to study variations in representations within and between subjects, as it is difficult for one or even a handful of subjects to provide enough data to learn representations for more than a trivial number of items.

Given the apparent requirement for large dataset sizes, it would be highly beneficial to develop more efficient representation learning methods. One possibility is to jointly model two pieces of information that are automatically collected in all behavioral tasks: choice and reaction time. Each of these has been used separately for representation learning: choices in the odd-one-out task have been modeled with the softmax choice rule (Hebart et al., 2020), and confusion matrices of reaction times in same-different judgments have been subjected to multidimensional scaling (Young, 1970). Jointly modeling these could combine potentially distinct sources of information in each, and therefore lead to more efficient representation learning than is possible with modeling choice or reaction time alone.

Interestingly, the potential of jointly modeling choice and reaction time for representation learning was recognized decades ago by Takane and Sergent (1983), who developed a joint model of choice and reaction time in same-different judgments, MAXRT. However, evaluations of this model have been lacking. Takane and Sergent (1983) fit their model to real data but did not compare it to modeling choice or reaction time alone. Storms and Delbeke (1992) fit MAXRT to real data to estimate item representations, $\hat{x}$, and MAXRT model parameters; used these estimated representations and model parameters to simulate new choices and reaction times; and then found MAXRT better recovered the item representations, $\hat{x}$, than a choice-only model, MAXSD. While these results are encouraging, this approach is essentially a parameter recovery simulation, since MAXRT was used to simulate the data on which it and MAXSD were fit and compared. Still needed is an empirical evaluation where choice-only and joint choice-RT models are compared in, e.g., their ability to predict independently established item representations or out of

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

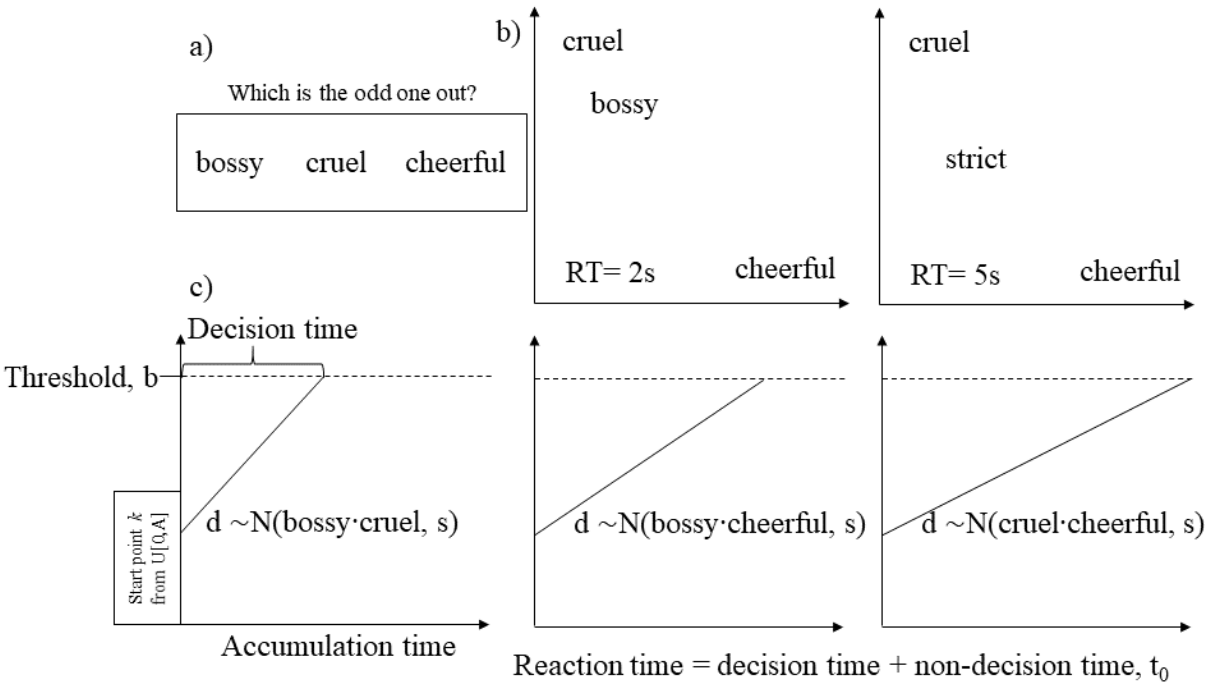

Figure 1: a) Odd-one-out task. b) 2-d item representations for two trials likely to generate short and long RT's. c) Overview of the Linear Ballistic Accumulator Model of the odd-one-out trial in (a).

sample choices based on estimated item representations.

Our contribution is therefore to provide such an empirical comparison between choice-alone and joint choice-RT models, in their efficiency of representation learning. As the joint models we develop are somewhat novel and distinct from MAXRT, we also conduct new parameter recovery simulations to first establish the in-principle value of joint modeling. In the rest of the paper, we describe the task we study, the models we consider, simulation and empirical studies of this task and model, and discussion of our results.

Odd-one-out task

Figure 1a illustrates the odd-one-out task which we study. In this task, a subject is given a set of three items, $\{i, j, k\}$, and must decide which item is least like the other two. This decision is equivalent to deciding which of the three possible pairs $\{(i, j), (i, k), (j, k)\}$ has the greatest overall similarity.

Models

Softmax choice model

Hebart et al. (2020) describe a model of choice in this task, where the probability of choosing pair (i, j) depends on the softmax choice rule

$$p(i,j) = \frac{\exp(\mathbf{x}_i \mathbf{x}_j)}{\exp(\mathbf{x}_i \mathbf{x}_j) + \exp(\mathbf{x}_i \mathbf{x}_k) + \exp(\mathbf{x}_j \mathbf{x}_k)} \quad (1)$$

where $\mathbf{x_i}$ is a d-dimensional, real-valued vector representation of item i, and $\mathbf{x_i x_j}$ is the dot product of the representations for items i and j[1]. Hebart et al. fit this model to human choices between object photos, and found interpretable item dimensions in terms of animacy, color, shape, and more. These representations also accurately predicted choices for held out odd-one-out trials, via the same softmax model.

Linear Ballistic Accumulator

Joint models of choice and reaction time typically come from the family of Evidence Accumulation Models (Evans & Wagenmakers, 2020). Although many EAMs might be applicable to the odd-one-out task, we tested the Linear Ballistic Accumulator (LBA), since it has a closed-form likelihood function for choice tasks with three or more options (Brown & Heathcote, 2008) and is implemented in the *rtdists* R package (Singmann et al., 2016). Figure 1 illustrates this model.

In the LBA, N accumulators track the amount of evidence gathered for each of N options. On each trial, each accumulator begins with a starting amount of evidence k, which increases at drift rate d, up to threshold b. The first accumulator to reach threshold determines the choice, and the RT is determined by the amount of time this accumulator takes to reach threshold plus some extra constant time for non-decision processes, t_0. Trial-by-trial variation in the LBA standardly comes from two sources: k, which is sampled from a uniform distribution $U(0, A)$, and d, from a normal distribution $N(v, s)$[2].

[1] Westfall and Lee (2021) describe a similar model, with the addition of weights on the features/dimensions of $\mathbf{x}$, and a Luce rather than a softmax choice rule (Luce, 2005).

[2] In principle, A, b, t_0, and s can vary by accumulator, to accom-

In our odd-one-out task, we have three accumulators, one for each pair in $(i,j),(i,k),(j,k)$, with the mean drift rate for an accumulator equal to the dot product of $\mathbf{x_i}$ and $\mathbf{x_j}$:

$$v(i,j) = \mathbf{x}_i \mathbf{x}_j \tag{2}$$

Thus, a highly similar pair (i,j) will have a high $v(i,j)$, and its accumulator will be more likely to reach threshold before other accumulators, triggering a choice of (i,j) as the most similar pair (or equivalently, k as the odd one out). Likewise, highly similar pairs will tend to be chosen more quickly than less similar pairs.

Other similarity metrics besides dot product are possible. For example, Hebart et al. (2020) found similar results with Euclidean distance instead of dot product. However, some metrics introduce challenges for modeling: cosine introduces an identifiability problem (if all representations double in magnitude, cosines are unchanged), and using softmaxed dot products as mean drift rates (i.e., Equation 1) proved unrecoverable in repeated simulations. For these reasons, we only report results with dot product.[3]

Softmax plus difficulty

An alternative joint model of choice and reaction time is inspired by Ballard and McClure (2019), who modeled choice and reaction time in multi-armed bandit reinforcement learning. Their model's loss function includes a term for the softmax log likelihood of choices, and another term for the log likelihood of a linear model of log reaction times [4]. This linear model includes terms for the linear and quadratic effects of the absolute value of the difference in values between bandits, under the intuition that trials with bandits closer in value would be more difficult, and hence increase decision time. We defined a similar model based on choice difficulty measured with the Shannon entropy in the softmax probabilities, and the overall sum of the dot products:

$$log(rt) = \beta_0 + \beta_1 * H(\mathbf{p}) + \beta_2 * H(\mathbf{p})^2 + \beta_3 * \sum \mathbf{v} \tag{3}$$

where $\mathbf{p}$ is a vector of choice probabilities from softmax, $\mathbf{v}$ is a vector of dot products between the three items' representations, and entropy is defined as:

$$H(\mathbf{p}) = -\sum_i p_i \log p_i \tag{4}$$

Notice that these two measures capture choice difficulty in a slightly different way: $H(\mathbf{p})$ predicts choices are difficult and slow when all pairs are equally similar (e.g., three tools) *or* equally dissimilar (e.g., a tool, a vehicle, and a building). $\sum \mathbf{v}$ predicts the former choice set will be fast and easy, and the latter slow and difficult.

Parameter recovery simulations

To demonstrate that, in principle, joint modeling of choice and reaction time improves representation learning over modeling choice alone, we conducted parameter recovery simulations. In each simulation, we first sample 20 2-dimensional vector representations from the unit square to produce a (20, 2) matrix, $\mathbf{X}$. We then simulatd choices and reaction times on all possible odd-one-out trials with our LBA model[5]. We selected LBA parameters that matched the empirical data we describe later, in terms of reaction time distributions and rates of between-subject agreement on the odd one out for a given trial. These parameters are $A = 2$, $b = 6$, $t_0 = .2$, $s = .1$, which we associate with a simulated subject, S_1. To simulate individual differences which might complicate joint modeling, we we also add another, slower but more careful subject, S_2, with $A = 1.8$, $b = 12$, $t_0 = .6$, $s = .08$, who responds to a different set of trials from S_1. Since S_1 will generally have faster RT's than S_2, we 'demean' each subject's RTs as follows: we log transform all RT's; calculate each subject's mean, M_1 and M_2, and the grand mean, M_g; calculate subject level shifts $shift_1 = M_g - M_1$ and $shift_2 = M_g - M_2$; add each subject's shift to their log RTs; and finally exponentiate all log RT's back to the raw scale. We then fit all models on these data with maximum likelihood estimation via *scipy*'s minimize function (Virtanen et al., 2020). We evaluate model performance in two ways. First, we calculate accuracy in choosing the odd-one-out among all trials not randomly selected for training. That is, if 25% of trials are randomly selected training, we calculate accuracy on the remaining 75% (this means that when training on all trials, we can not calculate out of sample choice accuracy). Second, we calculate the Procrustes disparity between true representations, $\mathbf{X}$, and estimated representations, $\hat{\mathbf{X}}$. Procrustes analysis finds the rotation, reflections, and scaling of $\hat{\mathbf{X}}$, that minimizes its disparity with $\mathbf{X}$:

$$M^2 = \sum (\mathbf{X} - \hat{\mathbf{X}})^2 \tag{5}$$

which is just the sum of squared elementwise differences between the two matrices.

We conducted the above procedure 20 times, when sampling 6.25%, 12.5%, 25%, 50%, and 100% of the possible trials. Figure 2 shows model performance at each level of sampling. As can be seen, modeling choice and RT with LBA recovers item representations much more accurately than does modeling choice alone with softmax, but LBA's advantage in

modate, e.g., a left or right choice bias. For simplicity, we do not model this.

[3] In fitting all three models with dot product as the similarity metric, we arbitrarily fix one item's representation to improve identifiability, since (a) dot product is rotation invariant, (b) and drift rate is known to trade off with other parameters in LBA (Brown & Heathcote, 2008).

[4] In principle, a hyperparameter could be used to trade off the impact of the softmax and RT loss terms, but following Ballard and McClure (2019), we omit this (but we return to this issue in the discussion.)

[5] We could also have simulated choices and reaction times with our Softmax Plus Difficulty model, but omit this for simplicity. The point of parameter recovery simulations is merely to demonstrate that joint modeling of choice and reaction time leads to more accurate representation learning than modeling choice alone.

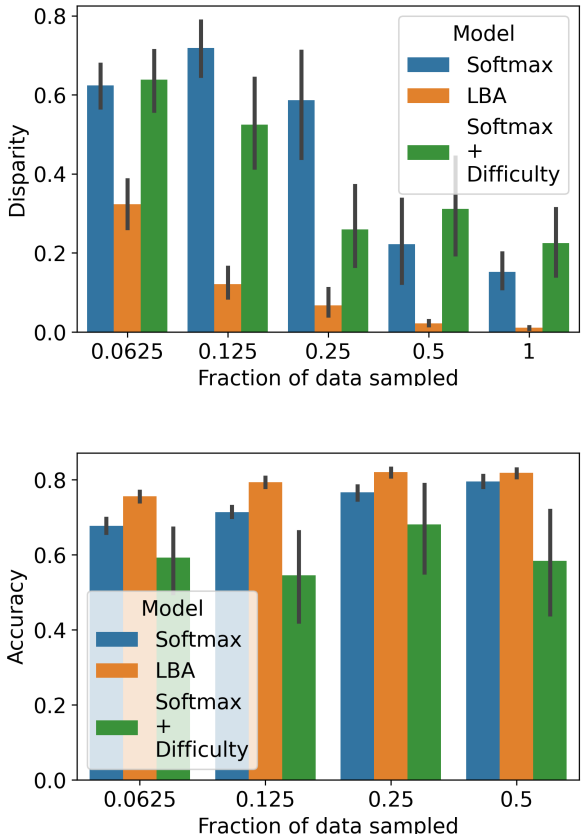

Figure 2: In each simulation, 2-dimensional item representations, $\mathbf{X}$, are generated; odd-one-out choices and RTs are simulated based on these item representations; and each model is fit to choice (and RT) to estimate the items' representations, $\hat{\mathbf{X}}$. Models are evaluated by calculating (top) disparity between $\mathbf{X}$ and $\hat{\mathbf{X}}$, or (bottom) accuracy in out-of-sample choices.

predicting out of sample choices is much weaker (an issue we return to when fitting empirical data). Perhaps surprisingly, Softmax Plus Difficulty generally performs no better than softmax alone.

Empirical tests

We now evaluate the usefulness of joint modeling of choice and reaction time for learning item representations in two real datasets.

Data

Hebart et al. (2020) Hebart et al. (2020) collected odd-one-out judgments for photos of 48 objects from different diverse categories, including foods, animals, tools, and vehicles. 100 subjects on Amazon Mechanical Turk completed 440 trials each, such that every triplet was judged by between two and eight subjects. Following Hebart et al., we filtered all subjects who chose one item (left, middle, or right) at least 40% of the time, or exhibited overly fast reaction times (25% or more responses $<$800 ms and 50% or more responses $>$1,100 ms). This removed about 10% of all data. We also removed every subject's first trial since this trial's RT was usually abnormally slow (subjects were calibrating to the study) and every trial $<$1000ms or $>$10,000ms, as these potentially reflected accidental presses, or cases where the subject was not attending to the choice for the duration (e.g., they may have taken a break). This removed another $\sim$5% of data.

One aspect of these real data that we did not model in simulations, is that subjects tend to get faster over time. To account for this and for individual differences simultaneously[6], we residualized log RT on trial number, a binary indicator for whether the trial was in the first 75 trials, the interaction of these two variables, and random intercepts for subjects. Thus, the fixed effects are equivalent to a piecewise regression split at the 75th trial. This was done as visual inspection of RT over time suggested that RT decreased most until this trial, and then leveled off. Residuals were added to the estimated overall intercept and then exponentiated to obtain raw RTs.

Trait words We collected a novel dataset of odd-one-out judgments on 60 trait words (e.g., bright, animated, critical) selected to adequately span the Big 5 personality factor space (Richie et al., 2020). 456 participants were recruited on Prolific Academic (all from the USA, fluent in English, with at least an 80% approval rate), and each completed 150 trait triplets interspersed with five catch trials where one trait word was replaced with a random noun (e.g., banana). Each trait triplet was tested twice, and subjects were free to take a break after trials 50 and 100. We applied the same filtering as above, and also removed subjects who failed more than one catch trial. This led to removal of 14% of all trials. We also applied detrending and demeaning pre-processing similar to that described above.

Results

Descriptive analyses of reaction time Before fitting our three models to these two datasets, we wished to verify that reaction times in these data are indeed reflective of similarity and choice difficulty. Thus, we measured the correlation between reaction time, and the two measures of choice difficulty utilized in the Softmax Plus Difficulty model: entropy and overall similarity.

For each pair of items in both datasets, we calculated the proportion of trials in which that pair was chosen as the most similar. This serves as a rough measure of similarity between a pair, and potentially as a measure of difficulty in a trial in which the pair appears. That is, a pair of items that are highly

[6]One necessary assumption for accounting for individual differences this way, is that each subject's trial set is more or less equal in difficulty or other aspects determining speed. If one subject had easier trials than another, then controlling for between subject variation would eliminate much of the trial-by-trial variance we wish to attribute to item representations. However, given that each subject does a random sample of 440 triplets, we are comfortable making this assumption. Similar assumptions hold in the simulations.

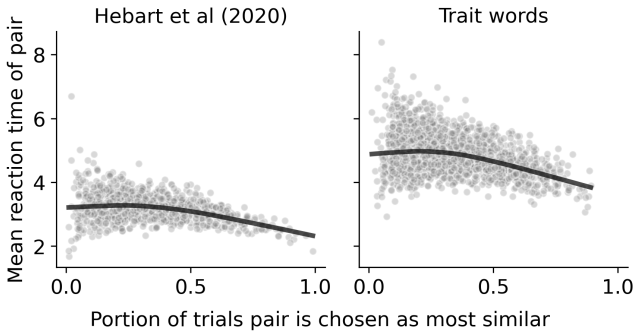

Figure 3: Percent of trials on which a pair of items is chosen as the most similar (an implicit measure of pair similarity and trial ease), versus mean reaction time on trials involving that pair. Reaction time is fastest on trials involving pairs frequently chosen as most similar (e.g., kind and nice).

similar (e.g., two images of tools, or near-synonymous traits like kind and nice), will be chosen at a high rate, and will generally be an easy, quick choice. Figure 3 plots each pair's choice proportion against the mean reaction time for trials involving that pair. Reaction times are indeed lowest for pairs that are chosen often. Reaction times are also highest around $p = .33$, consistent with the suggestion that choice is slowest and most difficult when entropy is maximal in the choice distribution ($\mathbf{p} = [.33, .33, .33]$). Reaction times are middling around $p = 0$ because choice entropy is maximized in this case when $\mathbf{p} = [0, .50, .50]$, which is less entropy than when $\mathbf{p} = [.33, .33, .33]$.

Indeed, in both datasets, the correlation between trial reaction time and the entropy in the choice proportions for each pair in the trial is around $r = .10$, $p < 10^{-60}$, suggesting reaction time is slower for more uncertain choices. The sum of the choice proportions for each pair in the trial – a measure of the overall similarity between all the items in the trial, and our second measure of trial difficulty – was also correlated with reaction time on a trial in both datasets, at $r = -.02$ in Hebart et al. (2020) and $r = -.04$ in the traits data, $p < 10^{-4}$ and $p < 10^{-17}$, respectively. This suggests reaction time is faster when the items are collectively more similar to each other, although the effects are very small.

Both of these RT effects are accounted for in the joint models: LBA accounts for the overall similarity effect, and Softmax Plus Difficulty accounts for both the entropy and overall similarity effects.

Cross-validation of model choice predictions Our first approach to empirical evaluation of models was to compare their ability to predict out-of-sample choices in 8-fold cross-validation. Figure 4 displays these results. We initially restricted model-fitting to two-dimensional representations, involving only the first 20 items in each dataset, to maintain tractability, as all analyses were implemented in CPU-based packages, on the first author's laptop, and used optimization methods that required the entire dataset fit in memory at once.

Figure 4: Cross-validation choice accuracy on empirical data. Contrary to expectations, for a given sample sparsity on both datasets, models accounting for choice and reaction time (LBA, Softmax + Difficulty), generally do not outperform models accounting for choice alone.

However, we later gained access to a high-performance computing cluster, and fit all 48 Hebart et al. (2020) images to four dimensional representations. Contrary to expectations, in both analyses involving 20 items fit to two dimensions, all models perform similarly for a given sample fraction on both datasets; in the analysis involving all 48 Hebart images fit to four dimensions, LBA performs worse than either softmax model, which perform similarly.

Comparing trait representations to Big Five Scores One possible objection to the previous analysis is that it inherently favors the softmax model, since softmax is designed to predict choice (and only choice), while the other two models are "forced" to accommodate an additional aspect of the data, reaction times (see Hawkins et al., 2014 for a similar argument in the context of best-worst scaling). Indeed, our parameter recovery simulations showed that, while LBA recovered item representations much more accurately than softmax, it had a much smaller advantage in predicting choices. Thus, in the absence of a real ground truth space to compare learned representations to (as is the case in simulations), we instead sought to compare each model's learned representations to independently obtained, well-established representations. For trait words, arguably these are subjects' ratings of how well different trait words describe themselves (Ashton, Lee, & Boies, 2015). Factor analysis of such ratings reveals familiar structure, with for example, a five factor solution yielding the Big 5: agreeableness, extraversion, etc. Each trait word's factor scores can thus be seen as a representation (with similar words having similar factor representations) to which we can compare representations learned from our odd-one-out data. For each model and training fold in our cross-validation analysis (at Fraction=1), we took the estimated trait word representations, and calculated their procrustes disparity with the two-factor solution of self-report trait ratings from Ashton et al. (2015). Softmax showed a lower disparity ($M = .67$, $SD = .03$) than LBA ($M = .69$, $SD = .03$) and Softmax Plus Difficulty (.70, $SD = .03$), although this difference was not significant, $F(1,21) = 2.45, p = .11$.

Discussion

Representation learning is a central problem in cognitive science, yet methods for representation learning are often data-hungry. Here, we explored the possibility that jointly modeling choice and reaction time in the odd-one-out task might improve data-efficiency in representation learning over modeling choice alone. Parameter recovery simulations were consistent with this, with a joint model based on the Linear Ballistic Accumulator outperforming a choice-only Softmax model. Empirical analyses of two datasets, one with images and one with words, also suggested that reaction time showed some correlation with item similarity and choice difficulty, but joint models of item representations were no better than choice-only models at predicting out of sample choices, or independently obtained factor scores for personality traits. Overall, these results present a somewhat negative picture regarding the utility of joint modeling of choice and reaction time for representation learning.

What accounts for the null results in our empirical model comparison? One possibility is that our joint models are misspecified. For example, mean RT is around 3s and 5s for Hebart et al. (2020) and our traits data, respectively, while Biele (2023) suggests 3s decisions are already stretching LBA's intended use case. Unsuitability of Softmax Plus Difficulty, however, would be surprising, given a similar model's success in a reinforcement learning task (Ballard & McClure, 2019). It could be that, as mentioned earlier, we need to fit a hyperparameter controlling the tradeoff between the softmax and difficulty terms in the Softmax Plus Difficulty model, although again, we point out that Ballard and McClure (2019) did not need to do this to find an advantage of joint modeling.

Another possibility may be that the relationship in these two datasets between RT and similarity/choice difficulty is too noisy (Takane & Sergent, 1983). Neither measure of choice difficulty correlated with RT above $r = .1$, while Ballard and McClure (2019) found average correlations of $r = .2$. This difference could be due to task differences, or because both our datasets were collected online while the data used by Ballard and McClure (2019) were collected in a lab (Wimmer, Braun, Daw, & Shohamy, 2014), which may have led to less noisy data. If noise is the culprit, it may be necessary to improve data collection methods and/or fit even larger datasets. While our dataset from Hebart et al. (2020) which exhausts all possible triplets among 48 objects contains more than 37,000 trials, the main dataset of Hebart et al. (2020) contains 1.46 million trials for 1,854 objects. Scaling to a dataset this size likely requires re-implementing all models in packages enabling GPU processing power and stochastic gradient descent.

Whatever the reasons, given the success of joint (choice, reaction time) modeling at improving latent parameter estimation in other domains (Zorowitz & Niv, 2023), we suggest that future research continue to explore the possibility of joint modeling in representation learning. Our data and code, available on OSF, hopefully assist next steps.

References

Ashton, M. C., Lee, K., & Boies, K. (2015). One-through six-component solutions from ratings on familiar english personality-descriptive adjectives. *Journal of Individual Differences*.

Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44.

Bhatia, S. (2017). The semantic representation of prejudice and stereotypes. *Cognition*, *164*, 46–60.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36.

Biele, G. (2023). Retrieved 2023-12-21, from https://discourse.mc-stan.org/t/warnings-when-using-rstan-for-linear-ballistic-accumulator-modelling/30217/7

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence accumulation models: current limitations and future directions. *Quantitative Methods for Psychology*, *16*(2), 73–90.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*, 381–386.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive science*, *38*(4), 701–735.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of spam: a fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256.

Kriegeskorte, N., & Mur, M. (2012). Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in psychology*, *3*, 245.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, and cognition*, *10*(1), 104.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Pitt, D. (2022). Mental Representation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University.

Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science*, *45*(8), e13030.

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2020). The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior research methods*, *52*, 1906–1928.

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), 50.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Cooper, L. A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychological science*, *3*(2), 97–104.

Singmann, H., Brown, S., Gretton, M., Heathcote, A., Voss, A., Voss, J., & Terry, A. (2016). rtdists: Response time distributions. *R package version 0.4-9. URL http://CRAN. R-project. org/package= rtdists*.

Steyvers, M. (2002). Multidimensional scaling. *Encyclopedia of cognitive science*, *1*.

Storms, G., & Delbeke, L. (1992). The irrelevance of distributional assumptions on reaction times in multidimensional scaling of same/different judgment tasks. *Psychometrika*, *57*, 599–614.

Takane, Y., & Sergent, J. (1983). Multidimensional scaling models for reaction times and same-different judgments. *Psychometrika*, *48*(3), 393–423.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Westfall, H. A., & Lee, M. D. (2021). A model-based analysis of the impairment of semantic memory. *Psychonomic Bulletin & Review*, *28*, 1484–1494.

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *Journal of Neuroscience*, *34*(45), 14901–14912.

Young, F. W. (1970). Nonmetric scaling of line lengths using latencies, similarity, and same-different judgments. *Perception & Psychophysics*, *8*(5), 363–369.

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.