You Take the High Road, and I'll Take the Low Road: Evaluating the Topographical Consistency of Cognitive Models

Sabina J. Sloman (SSLOMAN@Andrew.Cmu.Edu)
Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Daniel Oppenheimer (DOPPENH1@Andrew.Cmu.Edu)
Department of Social and Decision Sciences and Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

We present a novel framework for assessing the fit of cognitive models. Using this framework, we highlight limitations of existing methods of model evaluation, and derive new approaches to validating cognitive models. Tests of *topographical consistency* emphasize how a model's structure constrains behavior on *pairs* of coupled stimuli, even when point predictions on individual stimuli depend on estimates of the model's free parameters. By carefully selecting these coupled stimuli such that they follow the distinct topography of the model, researchers can overcome some limitations of existing methods. Finally, we provide a proof-of-concept example of how to use our approach to assess a model of multi-alternative, multi-attribute choice.

Keywords: model comparison; experimental design; decision-making

Cognitive scientists are often faced with the task of selecting between two or more models, yet the field is still rife with disagreement about the best way to evaluate candidate models (Busemeyer & Diederich, 2010; Lee et al., 2019; Myung & Pitt, 2018). Existing methods are typically dichotomized into two general classes: *qualitative* and *quantitative* model comparison techniques (Busemeyer & Diederich, 2010).

Testing a model qualitatively requires specifying a pattern of behavior that the model entails. As opposed to point predictions, qualitative predictions anticipate the *direction* of a behavioral trend, such as a preference reversal, a difference in performance on two tasks, or a decelerating perceptual curve.

Importantly, qualitative predictions are usually invariant to the specific parameter settings of a model. While qualitative tests are powerful tests of whether a model captures broad regularities, they are limited to cases in which the model's predictions *are* invariant to the specific parameter settings. Often, it's difficult to abstract general predictions from the space of all possible parameter combinations (Yechiam & Busemeyer, 2008).

Quantitative model comparison techniques estimate parameter values and use these estimates to make specific point predictions. Usually, models are selected on the basis of some fit statistic: an educated guess about how well the model will predict out-of-sample behavior. While quantitative methods facilitate comparison between much more complex models, this usually comes at the cost of interpretability: While qualitative tests demand specification of an anticipated behavioral pattern, quantitative tests impose no such constraint on the researcher. When relying on only quantitative techniques, it's often difficult to understand what the behavioral implications and diagnostic predictions of a model are (Birnbaum, 1974; Blaha, 2019; Navarro, Pitt, & Myung, 2004).

This paper establishes the theoretical grounding and presents an example application of an approach to model evaluation that leverages both the interpretability of qualitative tests, and the dependency of predictions on specific parameter values that quantitative tests allow for. *Topographical consistency* is the criterion that observed behavior falls along a model's specific topography in a pre-defined stimulus space. To drive the intuition for this method, consider a classic experimental paradigm from the decision sciences, which has been used to evaluate expected utility theory (EUT), a normative model of decision-making (Allais, 1953; Kahneman & Tversky, 1979): A participant is presented with a choice between two gambles. For any individual choice considered in isolation, EUT imposes no constraints on which of the two gambles the researcher can expect the participant to choose. However, assuming EUT implies that once the participant's choice has been observed, aspects of their stable utility function have been elicited, which constrains the space of choices the researcher can expect the participant to make next. The test of EUT (which is usually failed) is whether the theory allows the researcher to successfully predict patterns of choices across the stimulus space—what we refer to as the *topographical structure* of EUT. The method we propose essentially consists of generalizing this approach beyond utility functions. We demonstrate how the constraints imposed by the parameters of any model can be used to generate hypotheses about observed patterns of behavior on coupled stimuli.

The next section presents a novel graphical framework for representing the evaluation of a cognitive model. We use this framework to articulate the theoretical commitments made by existing methods, identify conditions that affect the diagnostic power of these methods, and highlight how the proposed method overcomes these limitations.

In 1979, George Box famously wrote "All models are wrong but some are useful" (Box, 1979). Here, we will use the term "useful" (in favor of "good" or "true") to denote a model that provides information about structural regularities in data, allowing the researcher to better, if not completely, anticipate patterns in their data.

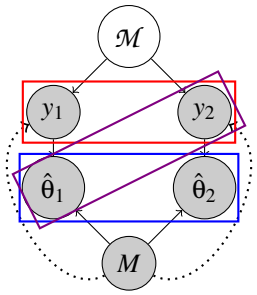

Figure 1: A graphical representation of the evaluation of model M. Shaded nodes indicate observed elements of the empirical world. See text for explanation of node labels. The purple box denotes the relation tested by traditional tests of out-of-sample generalization: the ability of $\hat{\theta}_1$ to predict y_2. The red and blue boxes denote the two other relations discussed in this paper: the ability of y_1 to predict y_2, and the ability of $\hat{\theta}_1$ to predict $\hat{\theta}_2$.

Model evaluation tests graphical relations

Researchers combine several observable elements of the empirical world to assess model fit: their data, the structure of the model, and the parameter values estimated by combining the model with their data. Figure 1 represents the relationships in this world as a graphical model. Each of these elements is represented as a node, while arrows, or *edges*, represent statistical dependencies between these elements. If there is a *path* between node A and node B—meaning that someone tracing edges from one node to the next would be able to reach node B after starting at node A—information from node A can be used to predict the state of node B (Shalizi, 2019).[1]

M represents the hypothesized model. $\mathcal{M}$ represents the unobservable cognitive processes of some group of participants. The researcher can collect one or more sets of observations of the behavior of these participants. Figure 1 shows the case where they collect two sets of observations, y_1 and y_2. The general framework does not restrict how y_1 and y_2 are collected. The researcher could randomly partition a set of stimuli into two sets, and refer to responses on one of them as y_1 and to responses on the others as y_2. In the following section, we explain the constraints our proposed method places on the environments in which to collect y_1 and y_2.

The researcher can estimate parameters of the hypothesized model M separately on y_1 and y_2. We refer to these parameterized versions of M as $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively.

Using this framework, we consider evaluating M's ability to fit y_1 and y_2 as testing the strength of the edges between M and y_1 and between M and y_2, shown in Figure 1 as dashed lines. If there is an edge between M and y_1 (y_2), i.e. M is useful, this implies that certain predictive relations hold. Modelers leverage this fact by testing these relations to assess model fit. For example, qualitative approaches to model evaluation test for the path $M \rightarrow y_1$ (or $M \rightarrow y_2$). If M is useful, the unparameterized M can be used to make predictions about the patterns in y_1.

Quantitative fit statistics generally try to approximate the ability of a fitted model to generalize, or to predict unseen observations (Gelman, Hwang, & Vehtari, 2014; Myung, Tang, & Pitt, 2009). In the context of Figure 1, these tests attempt to approximate the ability of $\hat{\theta}_1$ to predict y_2. If M is useful, there is a path between $\hat{\theta}_1$ and y_2, $\hat{\theta}_1 \leftarrow y_1 \leftarrow M \rightarrow y_2$, and $\hat{\theta}_1$ can be used to predict y_2.

However, if M is *not* useful—it contributes no information about y_1—there is still a path between $\hat{\theta}_1$ and y_2: $\hat{\theta}_1 \leftarrow y_1 \leftarrow \mathcal{M} \rightarrow y_2$. If y_1 and y_2 were collected in similar environments, $\hat{\theta}_1$ will contain information about y_2 insofar as the participants behave somewhat consistently. Broomell, Sloman, Blaha, and Chelen (2019) discuss one example of when relying on out-of-sample generalization can fail to be diagnostic of a path through a hypothesized model: Most parameterized models of risky decision-making will do extremely well at predicting participants' choices between almost all possible pairs of monetary gambles. This is not helpful in evaluating the relative usefulness of any of these models, but merely reflects the fact that for two randomly-chosen objects, it is usually the case that one is so much more valuable than the other that all reasonable models will make the same prediction. In general, selecting among models and among parameter values requires careful specification of the environments in which y_1 and y_2 are collected (Birnbaum, 1974; Broomell et al., 2019).

Qualitative and quantitative approaches to model evaluation each test one of the predictive relations in Figure 1 implied by M being useful. The graphical framework allows us to identify other paths M being useful implies. In the following sections, we unpack two other predictive relations modelers can test for to evaluate M.

As discussed just above, another contribution of the graphical framework is to illustrate that predictive relations can exist between behavioral patterns and parameter estimates even if the hypothesized model is not useful. Robust approaches to evaluating M test graphical relations that hold if—but *only if*—M is useful. We center discussion of our methods on explanation of how cognitive psychologists can implement them in a way that uniquely identifies paths through M. The crucial step is to collect y_1 and y_2 in *topographically distinct* environments: environments between which the model predicts systematically different behavior.

Relation #1: y_1 predicts y_2

Figure 1 shows that if M is useful, there is path from y_1 to y_2, $y_1 \leftarrow M \rightarrow y_2$, and the patterns in y_1 can be used to predict the patterns in y_2. How can a researcher collect y_1 and y_2 in a way that ensures that the information they gain about y_2 from y_1 flows through M, rather than through unobserved aspects of $\mathcal{M}$?

[1] There are exceptions to this general rule. For example, the path $M \rightarrow \hat{\theta}_1 \leftarrow y_1$ does not imply that M can be used to predict y_1 without taking $\hat{\theta}_1$ into account. See Shalizi (2019) for further discussion.

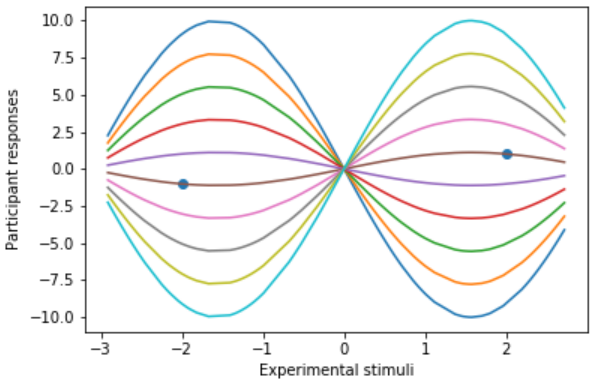

Figure 2: Parameter contours of the sine model (see text for details). The unparameterized sine model could take an infinite number of different shapes, and therefore does not constrain observed responses across much of the domain. The blue dots indicate observed responses that fall along a single contour. Under the sine model, if a participant responds -1 when presented with a stimulus value of -2, this implies the same participant would respond 1 when presented with a stimulus value of 2.

Consider a simple example: A researcher wants to test the usefulness of a hypothesized sine model: $y = \theta sin(x) + \varepsilon$. x represents manipulable aspects of possible experimental stimuli (e.g. light intensity, dollar value or time delay). y represents participants' responses to these stimuli, and ε represents random noise in these responses. θ is a free model parameter, unknown to the researcher. The sine model therefore makes a claim about the functional relationship between experimental stimuli and a participant's responses. Figure 2 shows possible relationships between x and y that would be consistent with the researcher's hypothesis. Because the researcher does not know any participant's θ *a priori*, they cannot constrain the set of possible responses. For example, the unparameterized sine model places no constraints on the participant's response when presented with a stimulus with value -2.

However, the model constrains *pairs* of observed data points. The researcher can exploit the structure of the sine model to make predictions about how a participant's responses to different stimuli will tend to "pivot" together.[2] For example, if the researcher observes a response of -1 when the participant is presented with stimulus value -2, they know that if the model is useful, the participant's response on 2 will be approximately 1. Similarly, if they observe that the participant's response on -2 is -10, they know to look for a response of approximately 10 when the participant encounters 2.

We call the topographical space created by iterating over these combinations of possible responses, shown in Figure 2, *parameter contours*. We think of these contours as a model's fingerprint. In effect, our approach consists of identifying observations that lie on the same contour. Tests of *topographical consistency* do not constrain which contour a participant is on, but rather evaluate whether the model's topography keeps the observations from each individual participant at the same "level." While one participant may take the "high road" and another the "low road," a useful model can predict how both of them will travel through the stimulus space.

Relation #2: $\hat{\theta}_1$ predicts $\hat{\theta}_2$

If M is useful, there are two paths from $\hat{\theta}_1$ (the parameterized version of M estimated using y_1) to $\hat{\theta}_2$ (the parameterized version of M estimated using y_2) through M: $\hat{\theta}_1 \leftarrow y_1 \leftarrow M \rightarrow y_2 \rightarrow \hat{\theta}_2$ and $\hat{\theta}_1 \leftarrow M \rightarrow \hat{\theta}_2$. So if M is useful, the researcher should gain information about $\hat{\theta}_2$ from $\hat{\theta}_1$. The predictive relationship is simple: The criterion described in the previous section tests whether the model generalizes the parameter estimate from a participant's behavior in one part of the stimulus space to their behavior in another. For the same reason the model predicts a participant's behavior will remain on the same parameter contour, it expects $\hat{\theta}_2$ to closely approximate $\hat{\theta}_1$.

Yechiam and Busemeyer (2008) refer to this implicative relationship as *individual parameter consistency*. We add to their criteria for model evaluation by highlighting the importance of testing the generalization of parameter estimates across data sets customized to the topographical structure of the model under investigation.

Regardless of whether or not M is useful, there is a path from $\hat{\theta}_1$ to $\hat{\theta}_2$: $\hat{\theta}_1 \leftarrow y_1 \leftarrow \mathcal{M} \rightarrow y_2 \rightarrow \hat{\theta}_2$. As above, careful stimulus selection is required to demonstrate that correspondence between $\hat{\theta}_1$ and $\hat{\theta}_2$ reflect M. To illustrate this, consider the researcher above, who wants to test the fit of the sine model to a given participant. Figure 3 shows two possible ways they could partition data from this participant. In the lefthand panel, the data set is partitioned randomly. As parameter consistency would predict if this model were useful, the two fitted sine curves are close together.

However, for comparison the researcher also tests the parameter consistency of a linear model. As shown in the lefthand panel of Figure 3, the fitted linear models also closely resemble each other. When y_1 and y_2 are collected in the same environments, parameter consistency is not diagnostic of either model.

Compare this to the partition shown in the righthand panel of Figure 3. Here, the researcher has selected topographically distinct regions of the sine model: Figure 2 shows that the sine model predicts that all participants will exhibit qualitatively different behavior when the stimulus value is greater than 0 than when it is less than 0.

When the data is partitioned into these topographically distinct regions, parameter consistency becomes diagnostic of the most useful model. While the estimated sine curves re-

[2]Our method exploits the assumption pervasive in the literature that parameter values are stable at the individual level (Glöckner & Pachur, 2012; Yechiam & Busemeyer, 2008).

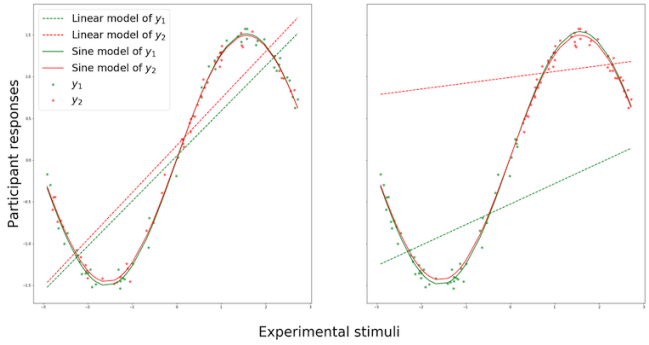

Figure 3: The sine and linear model fit to two different partitions of simulated data into y_1 and y_2 (the sine model is the data-generating model): a random partition (lefthand panel) and a partition based on the topographically distinct regions of the sine model's contours (righthand panel). Both panels show the best-fitting parameterized versions of the sine and linear models, fit separately to each partition.

main close together, demonstrating parameter consistency, the estimated linear curves have been pulled apart. This analysis exhibits clear support for the sine model (which is, in fact, the data-generating model).

In summary, tests of *topographical consistency* consist of identifying the topographical structure of a given model in a space of possible experimental stimuli, and testing whether participants' behavior follows the contours of this topography.

Application

To illustrate how to apply our approach, we develop a test of the Voting Agent Model of Preferences (VAMP) a one-parameter model of multi-alternative, multi-attribute choice (Bergner, Oppenheimer, & Detre, 2019). All data presented in this section are simulated, and results are intended as a proof of concept, rather than a true test of VAMP.

Multi-attribute choice refers to choice contexts where each option is defined by its value on the same set of attributes. For example, a decision-maker might be asked to imagine they were choosing between shoes, and given a numerical representation of each option's level of style and comfort (Bergner et al., 2019). Decision scientists have observed that participants make systematic yet anomalous choices when choosing between three options that relate to each other in certain ways (Roe, Busemeyer, & Townsend, 2001). There are dozens of models that attempt to account for these phenomena, including VAMP. For a review of other models, see Turner, Schley, Muller, and Tsetsos (2018) or Evans, Holmes, and Trueblood (2019).

VAMP has one parameter, typically referred to as k. For details on the model formulation, interested readers can consult Bergner et al. (2019).

Drawing contour plot. To generate the parameter contours of VAMP, we

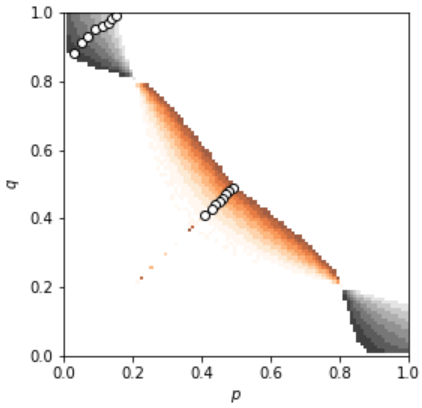

Figure 4: Parameter contours of VAMP. While in Figure 2 the stimulus space is one-dimensional, here it is two-dimensional (defined by both p and q, the attributes of option C). The two colors (gray and orange) denote topographically distinct regions of the stimulus space. The shading indicates the divergent predictions of adjacent values of k (see text for details). Dots denote selected stimuli.

1. Define constraints on our experimental stimuli. Here, we constrained a choice set to consist of three choices: $A = [.2, .8]$, $B = [.8, .2]$ and $C = [p, q]$ where $0 \leq p, q \leq 1$. This results in a two-dimensional stimulus space, with one dimension defined by the value of p, and the other defined by the value of q.

2. Segment the parameter space into discretized regions, in this case $\mathbf{k} = [.1, .15, .2, .25, .3, .35, .4, .45, .5]$. This range was theoretically motivated, as it approximates the range of k for which VAMP exhibits common decision anomalies (Bergner et al., 2019).

3. For each parameter value, compute the model's predictions on each element of the stimulus space. VAMP's predictions consist of a choice of one of the three options A, B or C.[3]

4. For each pair of adjacent parameter values (e.g. .1 and .15, .15 and .2, etc.), identify the sliver of the stimulus space on which the two parameter values make different predictions.

5. Stack these slivers to create a contour plot. VAMP's contours are illustrated in Figure 4. The shading indicates the divergent predictions of adjacent values of k. Darker colors indicate the regions that distinguish smaller values of k (e.g. the black/dark orange contour denotes the region that distinguishes $k = .1$ from $k = .15$).

$\mathbf{y_1}$ predicts $\mathbf{y_2}$. Figure 4 shows that different values of k make distinct predictions in the corner and center regions of the stimulus space. These regions are colored gray and orange in

[3]The version of VAMP we used for this example is probabilistic. These predictions refer to the maximizing choice.

Figure 4, respectively. Importantly, these distinct choice predictions are not arbitrary: They correspond to *shifts* in choice predictions. For all C defined by the corner (gray) regions of the space, VAMP predicts that decision-makers with $k = .1$ will select B. However, VAMP predicts that decision-makers with $k = .15$ will select A when the choice set includes a C that falls in the darkest contour, and B for all other stimuli in the region. Similarly, decision-makers with $k = .2$ are expected to select A when C falls in the bottom two contours, and B otherwise. A similar shift occurs in the center (orange) region of the space. For all C defined by this region, decision-makers with $k = .1$ are expected to select either A or B with equal probability. However, decision-makers with $k = .15$ are expected to select C when it falls in the darkest orange contour, and A or B otherwise. Decision-makers with $k = .2$ will select C in the top two contours, and A or B otherwise.

To generate y_1 and y_2, we partition the stimulus space across these *topographically distinct* regions. In exactly the same way the trough and peak of the sine wave are characteristic of the sine model, the gray and orange regions are characteristic of VAMP. If a decision-maker's behavior in the gray region (y_1) allows us to predict their behavior in the orange region (y_2), we consider this support for VAMP.

The dots in Figure 4 show the selected stimuli. These stimuli allow us to ground a test of VAMP in interpretable predictions such as "If decision-makers are more likely to select option B on stimulus [.09, .95], they should be more likely to select either A or B on stimulus [.46, .46]."[4]

Figure 5 illustrates the predicted behavioral pattern for $.2 < k < .25$. The coloring of the stimulus space denotes the maximizing choice for decision-makers with $.2 < k < .25$. The white strip denotes the corresponding contour, the region of the stimulus space for which decision-makers with $k = .2$ and $k = .25$ will disagree on the maximizing choice.

We simulated the choices of decision-makers with k uniformly distributed between .1 and .5. These decision-makers selected probabilistically between A, B and all C pictured in Figures 4 and 5. We then estimated each simulated decision-maker's k *only on the basis of their choices on stimuli in the corner (gray) region*.

The coloring of the dots indicates the choice share of the eight agents whose k was estimated to be between .2 and .25. In aggregate, there is a "switch point" exactly where the model predicts. Moreover, this information allows the model to predict where these decision-makers' choices will "switch" in a topographically distinct part of the stimulus space, the center region.

Importantly, we do not expect the point at which agents' choice switches to be the same for other values of k. Our predictions are not contingent on the exact location of the switch point. Rather, this method makes a prediction about the *direction* of movement of the decision-makers' choices: As choice shares tend to shift in one region of the space (y_1), this gives us information about how they will shift in another region of the space (y_2).

[4]Note that the selected stimuli not only span regions of the contour space, but span all contours. This was not strictly necessary for interpretable evaluation of VAMP, but constitutes a more stringent test of the model than selecting stimuli concentrated in fewer contours. See also Somerville (2019) for application of a similar approach used to recover the parameters of decision-making models.

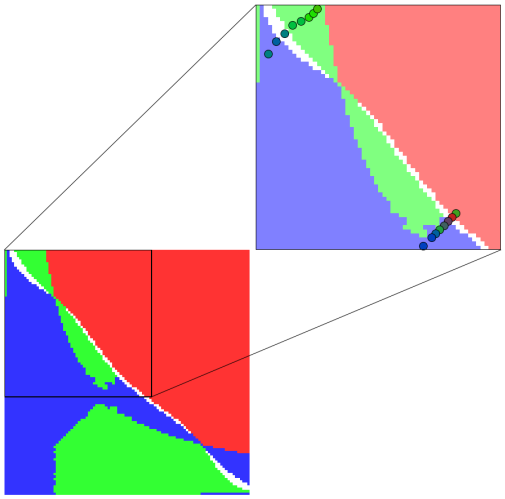

Figure 5: An illustration of the pattern of behavior expected if VAMP is useful. The coloring of the stimulus space denotes the maximizing choice for decision-makers with $.2 < k < .25$ (blue corresponds to A, green corresponds to B and red correspond to C). The white strip denotes the region of the stimulus space for which decision-makers with $k = .2$ and $k = .25$ will disagree on the maximizing choice. The color of the dots denotes the choice shares of simulated decision-makers whose k was estimated to be between .2 and .25 (the color correspondences are the same as above).

$\hat{\theta}_1$ predicts $\hat{\theta}_2$. The red points in Figure 6 show the relation between parameter estimates using data from one region ($\hat{\theta}_1$) and using data from another, topographically distinct region ($\hat{\theta}_2$). We compared these correspondences to those for a competing model: a slightly modified version of the pairwise normalization model (PN) proposed by Landry and Webb (2019). PN also has a single parameter, σ.

Using the same simulation approach described in the previous section, where VAMP is the known data-generating model, we used maximum likelihood to estimate a value of k and σ for each simulated decision-maker. Correlations between $\hat{k}_1$ and $\hat{k}_2$ are generally higher, although it is worth noting that as choice becomes less and less deterministic, the difference between the correlations reduces.

The blue points in Figure 6 show the same analysis, but with the stimuli shuffled, so they do not correspond to topographically distinct regions of the stimulus space. In this analysis, the correlations of VAMP's parameter become con-

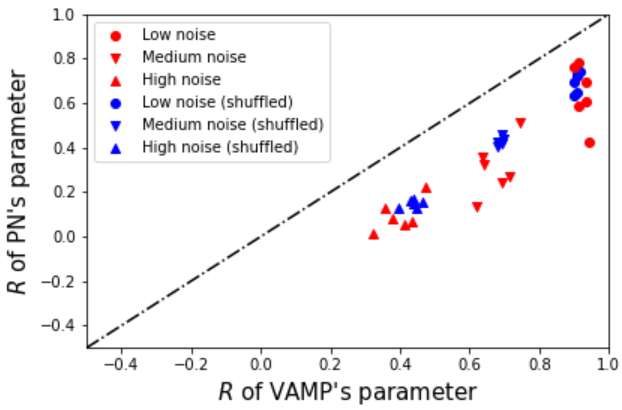

Figure 6: Correlations between parameter estimates for VAMP and PN, where VAMP is the generating model. x-axis: The Pearson's R between estimates of k on data generated from two non-overlapping sets of stimuli. y-axis: The Pearson's R between estimates of σ on the same data. Stimuli are partitioned into four sets; the partition differs for each color. Red: Stimuli are partitioned to fall in topographically distinct regions of the stimulus space. Two of these regions (the gray and orange regions shown in Figure 4) are discussed in the text. Blue: Stimuli are randomly assigned to a set. Each dot corresponds to a pair of stimulus sets and a value of noise, or non-determinism, in the simulated data. Values are mean correlations across 20 simulated datasets, each with an $n = 100$.

centrated around the center of the corresponding red points. However, the correlations of PN's parameter are reliably at the upper bound of the correlations exhibited when the stimuli are partitioned across VAMP's topographically distinct regions. In other words, most of the topographically distinct partitions succeed at better distinguishing VAMP from PN than the randomly-assigned partitions. This result highlights the importance of partitioning the stimulus set in a way that diagnoses paths through M.

Discussion

In this paper we have suggested *topographical consistency* as an evaluation criterion for cognitive models. To satisfy this criterion, a modeler commits to a prediction about how a participant's behavior in one part of the stimulus space varies with their behavior in a topographically distinct region of the stimulus space. Individual parameter consistency (Yechiam & Busemeyer, 2008) across these regions is an implication of the modeler's predictions being born out. We have not proposed a formal hypothesis test, made any suggestions about how dissimilar the modeler's observations should be from their predictions before the model is rejected, or discussed how to quantify this similarity. Formalizing these practical considerations is an area for future work; ultimately, though, the answers will likely depend on, among other things, in what way the modeler intends their model to be useful.

We applied the proposed approach to a one-parameter nonlinear model of multi-attribute decision-making. Extending our approach to models with more than one parameter is another avenue for future work. In general, it will require well-informed assumptions about the joint distributions of the parameters of the specific model being evaluated. These could be informed by analytical results or by unsupervised dimensionality reduction analyses of existing data sets (e.g. Yechiam and Busemeyer (2008)).

Topographical consistency reflects whether or not the observed patterns are *possible* under the hypothesized model. However, it does not directly tell the researcher how *likely* the hypothesized model considers the observed patterns to be. Many models assign a higher likelihood to some parameter values than others. As proposed, our approach imposes no penalty on a model if the distribution of estimated parameters violates such expectations. There exist many methods of quantifying and incorporating this consideration into the model evaluation process, e.g. parameter space partitioning (Pitt, Kim, Navarro, & Myung, 2006), representativeness analysis (Navarro et al., 2004), and Bayesian approaches (Farrell & Lewandowsky, 2018; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Combining our proposal to consider relationships between coupled pairs of stimuli that straddle topographical regions with these methods could lead to the development of even more robust approaches to model evaluation.

Further, core elements of the proposed approach can be incorporated into existing analyses of model fit. For example, we propose partitioning the stimulus space across topographically distinct regions as a principled method of stimulus selection. Modelers who use this approach to generate stimuli for their experiments could increase parameter identification (Somerville, 2019), and use the method of their choice to analyze the data from these experiments.

We motivated our approach with the graphical framework, and we see the criterion of topographical consistency as one instance of the approaches to model evaluation that can be derived from this framework. The framework makes explicit the convergent aims of existing model evaluation methods, establishes intuition for why tests of generalization should control for unmodeled sources of behavioral consistency, and can hopefully serve as a conceptual tool for modelers seeking to interrogate their hypotheses using a diverse set of methods. Tests of topographical consistency are not intended as a replacement for existing methods, and aspects of our approach can be incorporated in and contribute to these methods. Ideally, testing multiple criteria will help modelers better understand the characteristics of their models and the kinds of patterns each is most useful for predicting.

Acknowledgments

We thank Stephen Broomell, Leslie Blaha and Steven Sloman for helpful comments on a draft of this paper.

References

Allais, M. (1953, October). Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, *21*(4), 503. doi: 10.2307/1907921

Bergner, A. S., Oppenheimer, D. M., & Detre, G. (2019, November). VAMP (Voting Agent Model of Preferences): A computational model of individual multi-attribute choice. *Cognition*, *192*, 1–15. doi: 10.1016/j.cognition.2019.05.008

Birnbaum, M. H. (1974). Reply to the devil's advocates: Don't confound model testing and measurement. *Psychological Bulletin*, *81*(11), 854–859. doi: 10.1037/h0037132

Blaha, L. M. (2019, December). We Have Not Looked at Our Results Until We Have Displayed Them Effectively: A Comment on Robust Modeling in Cognitive Science. *Computational Brain & Behavior*, *2*(3-4), 247–250. doi: 10.1007/s42113-019-00059-6

Box, G. (1979). *Robustness in the Strategy of Scientific Model Building.* Mathematics Research Center, University of Wisconsin-Madison.

Broomell, S. B., Sloman, S. J., Blaha, L. M., & Chelen, J. (2019, December). Interpreting Model Comparison Requires Understanding Model-Stimulus Relationships. *Computational Brain & Behavior*, *2*(3-4), 233–238. doi: 10.1007/s42113-019-00052-z

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*. Thousand Oaks, California: SAGE Publications, Inc.

Evans, N. J., Holmes, W. R., & Trueblood, J. S. (2019, June). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychonomic Bulletin & Review*, *26*(3), 901–933. doi: 10.3758/s13423-018-1557-z

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. Cambridge, United Kingdom: Cambridge University Press.

Gelman, A., Hwang, J., & Vehtari, A. (2014, November). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi: 10.1007/s11222-013-9416-2

Glöckner, A., & Pachur, T. (2012, April). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*(1), 21–32. doi: 10.1016/j.cognition.2011.12.002

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291.

Landry, P., & Webb, R. (2019). Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2963863

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., ... Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*, 141–153. doi: 10.1007/s42113-019-00029-y

Myung, J. I., & Pitt, M. A. (2018, March). Model Comparison in Psychology. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–34). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9781119170174.epcn503

Myung, J. I., Tang, Y., & Pitt, M. A. (2009). Chapter 11 Evaluation and Comparison of Computational Models. In *Methods in Enzymology* (Vol. 454, pp. 287–304). Elsevier. doi: 10.1016/S0076-6879(08)03811-1

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004, August). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47–84. doi: 10.1016/j.cogpsych.2003.11.001

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006, January). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83. doi: 10.1037/0033-295X.113.1.57

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392. doi: 10.1037//0033-295X.108.2.370

Rouder, J. N., & Lu, J. (2005, August). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. doi: 10.3758/BF03196750

Shalizi, C. R. (2019). *Advanced Data Analysis from an Elementary Point of View.*

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008, December). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1248–1284. doi: 10.1080/03640210802414826

Somerville, J. (2019). *Choice-Set-Dependent Preferences in Consumer Choice: An Experimental Test.*

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018, April). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(3), 329–362. doi: 10.1037/rev0000089

Yechiam, E., & Busemeyer, J. R. (2008, May). Evaluating generalizability and parameter consistency in learning models. *Games and Economic Behavior*, *63*(1), 370–394. doi: 10.1016/j.geb.2007.08.011