Translating a Reinforcement Learning Task into a Computational Psychiatry Assay: Challenges and Strategies

Peter Hitchcock (pfh26@drexel.edu)
Clinical Psychology, Department of Psychology, Drexel University

Angela Radulescu (angelar@princeton.edu)
Department of Psychology, Princeton University

Yael Niv (yael@princeton.edu)
Department of Psychology, Princeton University

Chris R. Sims (chris.sims@drexel.edu)
Applied Cognitive & Brain Sciences, Department of Psychology, Drexel University

Abstract

Computational psychiatry applies advances from computational neuroscience to psychiatric disorders. A core aim is to develop tasks and modeling approaches that can advance clinical science. Special interest has centered on reinforcement learning (RL) tasks and models. However, laboratory tasks in general often have psychometric weaknesses and RL tasks pose special challenges. These challenges must be addressed if computational psychiatry is to capitalize on its promise of developing sensitive, replicable assays of cognitive function. Few resources identify these challenges and discuss strategies to mitigate them. Here, we first overview general psychometric challenges associated with laboratory tasks, as these may be unfamiliar to cognitive scientists. Next, we illustrate how these challenges interact with issues specific to RL tasks, in the context of presenting a case example of preparing an RL task for computational psychiatry. Throughout, we highlight how considering measurement issues prior to a clinical science study can inform study design.

Keywords: computational modeling; reinforcement learning; measurement; psychometrics; computational psychiatry

A core aim of the emerging field of computational psychiatry is to translate tasks and modeling approaches from computational neuroscience into sensitive assays that can advance clinical treatment, diagnosis, practice, and theory (Hitchcock, 2017; Redish & Gordon, 2016). New assays may advance clinical science by facilitating early illness detection, predicting illness progression, separating patients into subgroups, predicting type and extent of treatment indicated, and allowing measurement of the effects of emotion regulation strategies (Huys, Maia, & Frank, 2016).

The effort to develop laboratory tasks into assays has been ongoing for years, but the use of computational cognitive models that describe the trial-by-trial behavior of subjects (Daw, 2011) is newer to clinical science. In theory, parameters derived from these models should compactly describe individual or group differences by revealing aspects of cognitive processing that are obscured in behavioral measures (Huys et al., 2016). An especially promising domain in this regard is reinforcement learning (RL). RL refers to a broad class of trial-and-error learning tasks wherein learning is driven mainly by a scalar reinforcement signal (Sutton & Barto, 1998). Over the past twenty years, computational models of RL have grown in sophistication and maturity (O'Doherty, Cockburn, & Pauli, 2017). In addition, there has been a string of successful applications of RL modeling to clinical problems. These early successes may portend widespread use of RL assays in clinical science (Maia & Frank, 2011).

Yet the history of converting laboratory tasks to clinical assays suggests caution is warranted. Laboratory tasks tend to have substantial (and often underappreciated) psychometric weaknesses (Lilienfeld, 2014). Consider the example of the dot probe task, an attention paradigm introduced over 30 years ago (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & Van Ijzendoorn, 2007). By 2007, 35 clinical studies using the task had been conducted. A meta-analysis that year concluded the task reliably detects attention differences between anxious and non-anxious groups (Bar-Haim et al., 2007). Dozens of studies subsequently tested "modification" variants of the task (which aim to retrain attention) (Hallion & Ruscio, 2011). Yet recent meta-analyses suggest modification training produces very small effects and that extant modification studies evince publication bias (e.g., Heeren, Mogoaşe, Phillippbot, & McNally, 2015). These disappointing results prompted re-examination of the evidence for reliable, stable group differences per the original dot probe. Recent critiques, which have referenced a slew of null findings since 2007, concluded that the evidence for such differences is weak (Rodebaugh et al., 2016; Van Bockstaele, Verschuere, Tibboel, De Houwer, Crombez, & Koster, 2014).

What went wrong? It is noteworthy that, although researchers have been employing the original dot probe since the 1980s, the first examination of its test-retest reliability was not published until 2005 (Schmukle, 2005). That study and others (e.g., Price et al., 2015) found the dot probe exhibits close to 0 test-retest reliability when analyzed using standard methods. These results suggest it is not possible to extract stable measures of differences in attention using the standard versions/analyses of the task (Rodebaugh et al., 2016; Van Bockstaele et al., 2014).

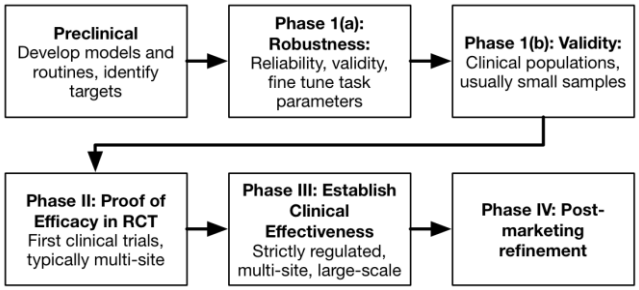

Figure 1. Pipeline for a computational psychiatry assay proposed by Paulus et al. (2016).

Developing Computational Psychiatry Tasks with Strong Psychometric Properties

The dot probe paradigm provides a cautionary tale about pushing too quickly from a lab paradigm to applied research. The computational psychiatry community can learn from this example. Fortunately, the community appears aware of the challenges posed by laboratory tasks. For instance, Paulus, Huys, and Maia (2016) proposed a pipeline (Figure 1) for turning a task into an assay that can ultimately be used for assessment or as a treatment target in randomized control trials (RCTs). The authors emphasize establishing psychometric properties early in the pipeline—before relying on the task as a primary measure in RCTs.

Yet researchers entering computational psychiatry from the cognitive sciences may be unfamiliar with how the psychometric challenges of laboratory tasks interact with clinical design issues. Thus, this paper offers an overview of the relevant issues. Specifically, the rest of paper is part theoretical overview and part annotated case example of preparing a specific RL task for use in clinical science. We begin with a theoretical issue that may be unfamiliar to many in cognitive science.

General Psychometric Challenges Associated with Laboratory Tasks. A general—and formidable—challenge for extrapolating from laboratory task behavior is that subjects naturally vary in the state that they are in (e.g., tired, distraught, cognitively taxed) when they arrive at the laboratory. A classic solution to this *random state variation* problem—a problem that confounded social and personality psychologists for decades (Kenrick & Funder, 1988)—is to assess the same subject at many time points and average over measurements. This approach can dramatically increase the convergent validity of lab tasks with self-report measures, presumably because an average over many time points yields a more stable, trait-like measure than one-time measurement (as the latter is often biased by state variation; Epstein, 1979).

However, assessing a single subject at many time points can be infeasible. First, much time is often needed to complete lab tasks, and thus repeating assessments on many occasions can substantially raise subject burden. Second, in some cases it is unrealistic to ask subjects to complete the task more than once. For example, a researcher may wish to examine how

Figure 2. The Dimensions Task (Niv et al., 2015), designed to investigate the role of attention in reinforcement learning.

depressive rumination—repetitive, negative, self-referential thinking—alters cognitive processing. This could be done by asking depressed subjects to complete a laboratory task while under the effects of a rumination induction. Many past studies have experimentally induced rumination in this way but, as far as we are aware, none has asked subjects to ruminate on more than one occasion. Indeed, it seems unreasonable and unrealistic to ask depressed subjects to undergo more than once a manipulation that—by design—provokes distress.

The effects of random state variation can be mitigated through study design. For example, a researcher investigating the effects of rumination on some task might ask subjects to perform the task once before and once while under the effects of rumination. Such a design should increase the ratio of systematic variability (variability due to induced rumination) to unsystematic variability (variability due to subjects being in different states when they enter the lab) because it delivers pre- as well as post-induction measures for each subject. A subject is unlikely to dramatically change the state she is in from pre- to post-induction (an unusually tired subject at baseline will likely remain so while under the effect of rumination induction). Thus the within-subject design controls for some of the unsystematic variability due to state variation. However, note that individual differences in susceptibility to the experimental perturbation (e.g., propensity to ruminate upon receiving the induction) will be affected by subjects' states. Thus this approach is helpful in minimizing noise but does not solve the random state variation problem.

The random state variation problem entails that a subject's parameter estimates in a laboratory measure will be corrupted by noise with respect to the subject's "true" parameter value, when the true parameter value is conceived of as a psychological variable akin to a trait. This noise will limit the predictive power of measures. Thus, when random state variation is expected (e.g., when a design only permits administering the task once or a few times), it is critical that the psychometric properties of the task are strong so that other sources of noise are minimized. For a more general discussion of how computational modeling may help remedy the random state variation problem, see Hitchcock (2017).

In the rest of this paper, we give a case example of preliminary efforts to establish the psychometric properties of a multidimensional RL task known as the *Dimensions Task* (Niv et al., 2015; Leong, Radulescu, Daniel, DeWonskin, & Niv, 2017; Radulescu et al., 2016). The task itself is not the paper's focus, but we briefly describe it, our approach to modeling it, and its promise for clinical science in the next

section. The description will make subsequent sections, on the task's measurement properties and their relation to modeling issues, easier to follow.

The Dimensions Task. Trial-and-error learning in the real world often requires learning about a small set of stimulus features embedded in a milieu of irrelevant stimuli. Imagine telling (what you hope is) an amusing story to a friend and attempting to learn about the effects of specific actions—dramatic pauses, rhetorical flourishes, funny faces, etc. Learning about the effect of these actions requires attending to just a few fleeting features on the face of and in the body language of your friend while ignoring many irrelevant features—pimples on your friend's forehead, your computer screen flickering behind you, your internal dialogue about what to say next, etc. (Niv et al., 2015).

The *Dimensions Task* was designed to study such a scenario where only some aspects of the task are relevant and most can be ignored, as is so often required in the real world. Briefly (see Niv et al., 2015 for details), on each trial of the task subjects must select one of three possible stimuli. Each stimulus is composed of 3 features defined on 3 stimulus dimensions (for example, color, shape, and pattern) (Figure 2). Subjects play a set of games that can vary in length from 15-30 trials. Within a game, features of only one dimension (e.g., color) determine the probability of reward. Within this *relevant* dimension, one *target* feature (e.g., red) leads to reward with 75% chance whereas the other 2 features in the dimension (e.g., yellow, green) lead to reward with 25% chance. The target feature and relevant dimension change every game. The start of a new game is signaled to subjects.

Computational Model. Previous work (e.g., Niv et al. 2015; Radulescu et al., 2016) tested various computational models designed to reproduce subjects' trial-by-trial behavior in the task and found that human behavior is well described by a *feature-level RL (fRL)+decay* model. The *fRL+decay* model maintains weights reflecting the values of each of the 9 features. It linearly sums these weights to calculate the estimated value of each (3-feature) stimulus

$$V(S) = \sum W(f) \qquad \forall f \in S \tag{1}$$

For example, the model's estimate of the value of yellow-waves-triangle in the above trial is equal to the sum of the weights of yellow, waves, and triangle. Once a reward is received (0 or 1 points), the weights of the 3 features of the selected stimulus are updated based on the discrepancy between the obtained reward, R_t, and the model's estimate of the chosen stimulus's value, $V(S_{Chosen})$, with update rate controlled by a learning rate free parameter, η

$$W^{new}(f) = W^{old}(f) + \eta[R_t - V(S_{chosen})] \qquad \forall f \in S_{chosen} \tag{2}$$

For the other 6 features on a trial—those comprising the 2 stimuli *not* selected—the model decays the associated weights with a second free parameter, d

$$W^{new}(f) = (1-d)W^{old}(f) \qquad \forall f \notin S_{chosen} \tag{3}$$

The decay parameter reflects the fact that subjects are selectively attending to (and learning about) few dimensions (Leong, Radulescu et al., 2017). The "forgetting" of the weights of unchosen features allows the model to "undo" learning about features not chosen on a trial.

Finally, the model assumes that the subject's probability of choosing each stimulus is proportional to the estimate of the value of the stimulus, as defined by a softmax equation with a third free parameter, β

$$p(\text{choose } S_i) \propto e^{\beta V(S_i)} \tag{4}$$

The model thus has three free parameters: softmax action selection noise β, learning rate η, and decay parameter d. See Niv et al. (2015) for more details.

Stage in the Assay Development Pipeline. With respect to Paulus et al.'s (2016) pipeline (Figure 1), most prior studies using the Dimensions Task and *fRL+decay* model fall into the Preclinical and Phase1a phases.

Notably, Radulescu et al. (2016, study 2) also provided a test of the task's promise for measuring group differences. Radulescu and colleagues found older adults were less accurate (p = .001, g = .94) than younger adults. These behavioral results appeared to derive in part from differences in the decay parameter (median = .52 v .42 for older vs. younger adults, respectively), implying that differences in this parameter may reflect meaningful differences in selective attention. These results suggest the task has promise as a sensitive measure of neuropsychological and clinical differences. Per Paulus et al.'s (2016) pipeline, this study marks the entrance into Phase 1b: examining clinical validity (see Radulescu et al., 2016 for discussion).

Although the task has promise as a computational psychiatry assay, a number of modeling and psychometric obstacles must first be overcome. In the following sections, we report on efforts to explore the properties of the *Dimensions Task* and *fRL+decay* model using two previously collected datasets. The results have implications for the use of the *Dimensions Task* in computational psychiatry and thus are of specific interest to researchers interested in the construct of attention learning in computational psychiatry. But the more general interest aim of the following sections is to use this case study to illustrate some of the issues that arise in translating RL tasks to computational psychiatry.

Methods

Datasets are from Niv et al. (2015; hereafter **D1**) and Radulescu et al. (2016, study 2; hereafter **D2**).

Specifications. In D1 (N = 22), subjects played 500 trials (number of trials per game was drawn from a Uniform(15,25) distribution, for a total of M=22.27, SD=1.45 games per subject). In D2 (N = 54), subjects played ~1400 trials (M=46.43, SD=5.41 games; subjects stopped playing after exactly 40 min.; all games 30 trials).

Results

Parameter Identifiability. A challenge in fitting RL parameters to individual subject behavior is that parameters

can be coupled and thus not fully identifiable. In the *fRL+decay* model, equations 1–4 show that the role of each parameter depends on the settings of the other parameters. Specifically, the values of the stimuli in equation 4 (in which choice is governed by β) depend—via equation 1—on the weights of the chosen and non-chosen stimuli. Those weights are in turn respectively governed by the learning rate (η, equation 2) and decay rate (d, equation 3).

Coupling of the parameters modulating value estimation and choice is characteristic of many RL algorithms (Daw, 2011; Gershman, 2016). Coupling comes in two flavors: severe and moderate (Daw, 2011). Under severe coupling, parameters can *trade off*; for example, increases in one parameter can be perfectly compensated by decreases in another. As a result, parameter values may not—even in principle—be uniquely identifiable. Severe coupling can be tested for by repeatedly run an off-the-shelf optimizer from different initial parameter settings and checking whether optimization converges on the same estimates every time. If parameters are structurally coupled (i.e., there is no unique set of estimates), the optimizer will find different estimates on different runs, provided initializations allow the optimizer to cover sufficient territory in likelihood space. In D1 and D2, an optimizer repeatedly converged on the same parameter estimates, suggesting identifiability issues are not too severe to prevent finding a unique optimum.

However, there may still be more moderate identifiability issues. Intuitively, this is because maximum likelihood/maximum a posteriori (ML/MAP) estimates are tantamount to finding the highest point on the "hill" that defines the parameter surface in likelihood/posterior probability space. Yet they do not reveal the shape of the hill below: specifically, the shape of equal-likelihood ridges in the 3D likelihood space. If these ridges are diagonally shaped, they indicate covariance between the parameters. Intuitively, if changing a parameter in one direction (e.g., η from 0.08 to 0.1) can be compensated for by changing another (e.g., β from 6.2 to 5.1), with only miniscule changes in the likelihood, then one cannot safely draw conclusions from the point estimate of either parameter.

Identifiability and Computational Psychiatry. Identifiability poses a special challenge in the computational psychiatry domain, wherein the aim is often to derive parameters that can be used as predictors or outcome measures (Huys et al., 2016). Derived parameters whose point estimates have much uncertainty about them due to identifiability issues are unlikely to be useful for precision applications, such as prediction or diagnostic subtyping.

Probing Identifiability. A first helpful step for probing identifiability is to examine and visualize the Pearson correlations between pairs of estimates. Figure 3 plots point estimates for pairs of parameters in D1 and D2, with regression lines drawn to aid visualization.

Sets of parameters can fall along an elliptical contour in the likelihood space if there are identifiability issues, in which case the parameters will correlate. Thus, if parameter pairs closely correlate for most subjects in a dataset, this may

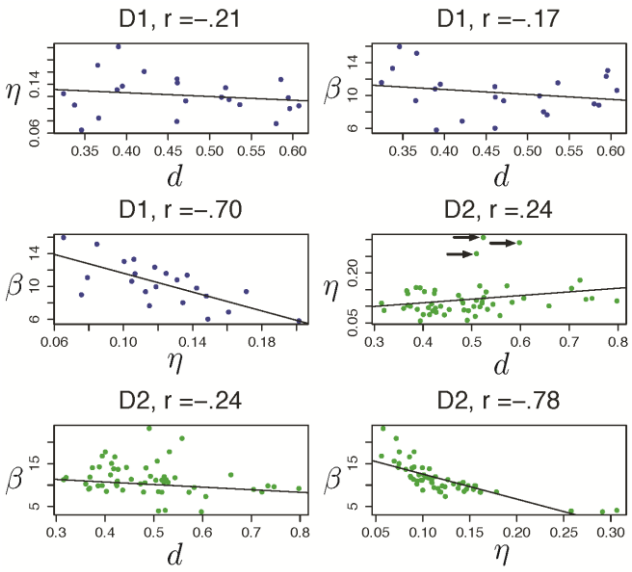

Figure 3. Parameter estimates in D1 (blue) and D2 (green).

indicate identifiability issues. However, correlations should only be a first step in checking for identifiability issues, for a couple reasons. First, to the extent that the parameters reflect meaningful psychological differences between individuals, we should expect they will correlate to some degree, because psychological variables often correlate within-subject (Lykken, 1968). Thus it can be difficult to determine whether correlations reflect modeling noise or true correlations between parameters. Second, correlations will not detect non-linear relationships between parameters or other subtle identifiability issues (Gershman, 2016). Still, correlations are easily interpretable and a good place to start.

Figure 3 shows that, in both D1 and D2, {d and β} and {d and η} modestly correlate whereas {η and β} strongly correlate. In particular, in D2, {η-β} estimates are nearly perfectly collinear for many subjects. Note also that, for all parameter pairs, the correlations are higher in D2, where there were more data, than in D1. In the test-retest reliability section below, we will present evidence suggesting that the parameter estimates may be more reliable in D2 than D1. Yet the higher correlations may also suggest more identifiability problems in D2. In fact, both possibilities—better parameter estimates and more identifiability problems—may be true. As noted, the equations in which the parameters are embedded dictate dependencies—and hence identifiability issues—between the parameters. If the true parameters are correlated, then when the model does a better job of recovering their values from noisy behavioral data, the observed data will also correlate more strongly. Thus, the increased correlations may actually be good news from a parameter recovery perspective. However, the {η-β} collinearity does mean we should not treat these variables as independent.

Diagnosing Issues with Model Fit. Plots also allow visualization of outlying values, which may reflect model fit issues for specific subjects. For example, the arrows in the second row, second column plot in Figure 3 point to subjects

with outlying η values. Outlying values might indicate *fRL+decay* does not well describe specific subjects' choices in some or all of the task. However, the points could also reflect important individual differences, so additional checks are necessary to make a differential diagnosis.

We do not delve into model fit issues for individual subjects (as such specifics would not generalize beyond the samples/data in D1 and D2), but offer some general guidelines for probing these issues. First, another useful diagnostic is to plot likelihoods for each potentially problematic subject. For example, Daw (2011) provides an example of a 2D heat map of likelihood values. A more quantitative assessment is the variance—covariance structure of parameter estimates; these structures can be examined by taking the inverse of the Hessian from optimization. On- and off- diagonal elements of H^{-1} respectively give the variance and covariance of parameters. Large values indicate poor parameter estimates (Daw, 2011). Finally, problematic subjects' behavioral (and physiological, if available) data can be checked to see if these data are informative about the source of outlying parameter values (e.g., if reaction times were recorded, it can be useful to check if a subject responded atypically quickly or slowly during a subset or all of the task).

Ultimately, if outlying parameter values for a subject do not appear to be due to individual differences, but rather to issues with model-fit, the researcher may wish to treat these parameters as missing: Parameter estimates derived from a model that poorly describes a subject's behavior are meaningless. However, such decisions should be made—then adhered to—before inferential statistics, to avoid the "garden of forking paths" (Gelman & Loken, 2013).

Subject-Specific Model Fit Issues and Computational Psychiatry. Our identification of apparent model fit issues among subjects illustrates the value of collecting data under different tasks specifications prior to attempting to develop a computational psychiatry assay. For instance, in the *Dimensions Task*, the presence of multiple individuals with apparently poor model fits suggests that some subjects in future clinical science designs will likely have missing data for model parameters (because, as noted, values from a model that poorly describes participant behavior should not be used). This is important information in the design phase of a clinical science study, as it may influence factors such as recruitment target, or collection of other data to aid estimation of anticipated missing values.

Test-retest reliability. As the cautionary tale of the dot probe task suggests, it is critical to establish the test-retest reliability of potential outcome measures. High test-retest reliability scores increase confidence that the measure is tapping a stable psychological construct (Hitchcock, Radulescu, Niv, & Sims, 2017). Establishing stability of a measurement is a prerequisite for computational psychiatry designs that seek to use the measure to assess the effects of some experimental perturbation or group or individual differences. Nevertheless, the basic requirement of establishing test-retest reliability goes unmet with striking frequency in laboratory tasks (Lilienfeld, 2014).

Dataset	d	η	β
D1	.17	.36	.71
D2	.68	.79	.69

Table 1. Intraclass correlation coefficients of parameters.

Table 1 presents test-retest reliability data for D1 and D2. These estimates were derived from splitting the data into approximately equal halves (specifically at the first game change after half of trials elapsed) and fitting the model to each (approximate) half. The test-retest reliabilities for {d and η} in D1 were quite low. This is likely because subjects only played 500 trials, and ~250—the approximate number of trials per half—may be too few trials to reliably estimate the parameters. In contrast, the D2 data suggest that ~700 trials allows for better parameter estimation, as reflected in the fact that test-retest scores for {d and η} are much higher.

Universal norms for intra-class correlation coefficients (ICCs) are arguably not justifiable (Weir, 2005) and at present there are no ICC benchmarks for RL tasks. But, in all domains, uncertainties around parameters increase as ICCs decrease (Weir, 2005). Thus, the above data are relevant to clinical science designs because they show how ICCs can increase with more data (see also Hitchcock et al., 2017). Gathering this information before designing a computational psychiatry assay is useful because computational psychiatry designs must often balance competing goals. On one hand, parameter estimates tend to improve with more trials. On the other, it may be infeasible to have subjects complete too long a task. For instance, individuals with certain disorders may fatigue easily. Experimental manipulations (e.g., rumination inductions) may also quickly dissipate. Test-retest reliability data can help negotiate the tradeoff between optimizing parameter estimates and keeping time on task feasible.

Conclusions

Computational psychiatry promises to improve measurement and refine theory in clinical science (Hitchcock, 2017). Ultimately it may advance understanding of psychiatric disorders (Redish & Gordon, 2016). Yet there are significant barriers to developing computational psychiatry assays. These barriers are diverse; hence this paper was part theoretical overview and part case study. The overview part of the paper first built motivation by discussing the dot probe paradigm, a case in which failure to attend to measurement issues in a laboratory task had disastrous results. Dozens of studies were conducted and vast resources were expended, over decades, before the poor properties of the task measures were realized. Next, we reviewed why laboratory tasks are so vulnerable to measurement issues: Task performance is often skewed by random state variation. That is, behavior collected only once or a few times from a single subject is often corrupted by situational factors. These review parts of the paper highlighted that minimizing noise in laboratory task measures is imperative. In the case study part of the paper, we overviewed modeling issues in RL tasks that can add

noise to parameter estimates, using two datasets for illustration. We concluded by presenting test-retest reliability data from the *Dimensions Task*, using this example to illustrate how time-on-task can improve reliability.

We should note that we have presented only some of the steps that should be taken when applying an RL task in clinical science. Other options include applying empirical priors (Gershman, 2016), using physiological data to aid parameter estimation (e.g., Leong, Radulescu, et al, 2017), and employing hierarchical modeling to weight parameter estimates by group statistics (Gelman & Hill, 2006), which can reduce the variance of parameter estimates (Daw, 2011). As computational psychiatry develops, we predict that psychometric, study design, and parameter estimation issues will come increasingly to the fore.

Acknowledgments

This work was supported by NSF research grant DRL-1560829 (CRS) and ARO grant W911NF-14-1-0101 (YN).

References

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psych. Bull.*, *133*(1), 1-24.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models (pp. 3-38). *Decision making, affect, and learning: Attention and performance XXIII*, *23*.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social psychology*, *37*(7), 1097-1126.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Department of Statistics, Columbia University, New York, NY.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *J. Mathematical Psychology*, *71*, 1-6.

Hallion, L. S., & Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological Bulletin*, *137*(6), 940-958.

Heeren, A., Mogoaşe, C., Philippot, P., & McNally, R. J. (2015). Attention bias modification for social anxiety: a systematic review and meta-analysis. *Clin. Psych. Rev.*, *40*, 76-90.

Hitchcock, P.F. (2017). Computational Modeling and Reform in Clinical Science. [preprint; osf.io/mvxkf] *OSF*.

Hitchcock, P.F., Radulescu, A., Niv, Y., Sims, C.R. (2017). Assessing the Potential of Computational Modeling in Clinical Science. In *The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision-Making.*

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neuro.*, *19*(3), 404-413.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*(1), 23-34.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451-463.

Lilienfeld, S. O. (2014). The Research Domain Criteria (RDoC): an analysis of methodological and conceptual challenges (pp. 13-14). *Beh. Res. and Ther.*, *62*, 129-139.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psych. Bull.*, *70*(3), 151-159.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154-162.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. of Neurosci.*, *35*(21), 8145-8157.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*, 73-100.

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 386-392.

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., ... & Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment*, *27*(2), 365-376.

Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and Aging*, *31*(7), 747-757.

Redish, A. D., & Gordon, J. A. (2016). *Computational Psychiatry: New Perspectives on Mental Illness*. Cambridge, MA: MIT Press.

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology*, *125*(6), 840-851.

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*(7), 595-605.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT press.

Van Bockstaele, B., Verschuere, B., Tibboel, H., De Houwer, J., Crombez, G., & Koster, E. H. (2014). A review of current evidence for the causal impact of attentional bias on fear and anxiety. *Psychological Bulletin*, *140*(3), 682-721.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The J. of Str. & Cond. Res.*, *19*(1), 231-240.