Simulating Category Learning and Set Shifting Deficits in Patients Weight-Restored from Anorexia Nervosa

J. Vincent Filoteo[a,2], Erick J. Paul[3], F. Gregory Ashby[4], Guido K.W. Frank[5], Sebastien Helie[6], Roxanne Rockwell[2], Amanda Bischoff-Grethe[2], Christina Wierenga[2], & Walter H. Kaye[2]

[1]Veterans Administration San Diego Healthcare System;
[2]Department of Psychiatry, University of California San Diego;
[3]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign
[4]Psychology Department, University of California Santa Barbara
[5]Departments of Psychiatry and Neuroscience, University of Colorado Anschutz Medical Campus
[6]Department of Psychological Sciences, Purdue University

Objective: To examine set shifting in a group of women previously diagnosed with anorexia nervosa (AN) who are now weight-restored (AN-WR) and a control group and then apply a biologically-based computational model (Competition between Verbal and Implicit Systems; COVIS) to simulate the pattern of category learning and set shifting performances observed in the AN-CW group. Method: Nineteen AN-WR women and 35 control women (CW) were administered an explicit category learning task that required the initial acquisition of a rule, and after a certain number of trials, a set shift following a rule change. COVIS was first fit to the behavioral results of the controls and then parameters of the model theoretically relevant to AN were altered to mimic the behavioral results. Results: Relative to CW, the AN-WR group displayed steeper learning curves (i.e., hyper learning) prior to the rule shift, but greater difficulty in learning the new categories after the rule shift (i.e., a deficit in set shifting). Hyper learning and set shifting deficits in the AN-CW group were not associated and demonstrated a different pattern of correlations with clinical measures. Hyper learning in the AN-WR group was simulated by increasing the model parameter that represents sensitivity to negative feedback (δ parameter), whereas the deficit in set shifting was simulated by altering the parameters that represent changes in rule selection and flexibility (λ and γ parameters, respectively), processes dependent on dopamine levels. Conclusions: These simulations suggest that multiple factors can impact category learning and set shifting in AN-WR individuals (e.g., alterations in sensitivity to negative feedback, rule selection deficits, and inflexibility) and provide an important starting point to further investigate this pervasive deficit in adult AN.

A consistent finding in the neuropsychology of eating disorders is a pervasive and persistent deficit in the ability to shift cognitive set. Set shifting is a cognitive concept that refers to the ability to switch tasks or change behavior in relation to changing rules. Set shifting is often evaluated by having a participant learn a particular rule using feedback and then switching the rule covertly after a certain number of correct responses. A set shifting deficit is observed when the participant fails to switch to the new rule but rather persists with the previously correct rule. Adult patients with Anorexia Nervosa (AN) are often impaired in making such set shifts, as demonstrated by a number of studies that found currently ill AN patients to be impaired on the Wisconsin Card Sorting Test (WCST), as well as other tasks that require set shifting (Roberts, Tchanturia, Stahl, Southgate, & Treasure, 2007; Roberts, Tchanturia, & Treasure, 2010; Shott et al., 2012; Steinglass, Walsh, & Stern, 2006; Tchanturia et al., 2011). Reduced set shifting abilities are also observed in unaffected relatives of AN patients (Roberts et al., 2010; Tenconi et al., 2010) and persist even after individuals with AN have restored their weight to normal levels (Danner et al., 2012; Roberts et al., 2010; Tchanturia, Morris, Surguladze, & Treasure, 2002; Tenconi et al., 2010). These findings are consistent with the clinical observation that, from a personality perspective, patients with AN tend to be rigid, nonflexible, harm avoidant and perfectionistic (Casper, Hedeker, & McClough, 1992; Merwin et al., 2011). The combination of a rigid personality style along with cognitive set shifting deficits has important implications given that these behaviors could lead to the development of the disease and impact the potential for recovery (Merwin et al., 2011; Roberts et al., 2010).

Despite the consistent finding of a set shifting deficit in patients with AN, few studies have offered specific insights into the nature and potential mechanisms of this deficit. One recent study (Zastrow et al., 2009) using

Dr. Filoteo's work was supported in part by a VA Merit Award. Dr. Ashby's work was supported in part by grants from the National Institute of Neurological Disorders and Stroke (P01NS044393) and from an award from the U.S. Army Research Office through the Institute for Collaborative Biotechnologies (W911NF-07-1-0072). Dr. Kaye's work was supported by grants from the National Institute of Mental Health (MH046001, MH042984, MH066122; MH001894 and MH092793), the Price Foundation, and the Davis/Wismer Foundation.

fMRI attempted to examine the neurobiological basis of set shifting deficits in currently ill AN patients. These investigators found few differences between AN patients and controls in functional brain activation on trials when a set shift was required. Thus, this study did not help elucidate the neural mechanisms that might be associated with set shifting deficits.

In the present study, we take a slightly different approach to better understand the set shifting deficits observed in AN. Here we examine set shifting in a group of participants who were previously diagnosed with AN but are now weight-restored (AN-WR) using a task on which we have recently demonstrated currently ill AN patients to be impaired (Shott et al., 2012). This task is a somewhat novel, but simple, category-learning task that requires the learning of changing rules, thereby emphasizing set-shifting abilities. The task has a set number of trials both prior to and following the rule change (set-shift), thereby allowing us to examine both the speed of acquiring the new rule as well as the ability to shift to a new rule. This is in contrast to other clinical measures that have been used in the past to examine set-shifting in AN, such as the WCST, where the number of trials prior to a shift requirement is not pre-determined, thereby allowing each participant to have a different amount of exposure to the initial rule. We focused on weight-restored participants to examine whether set-shifting deficits persisted once individuals with AN no longer met criteria for the disease and to rule out the possibility that any observed deficits were associated with the acute stages of the disease (e.g., current malnutrition).

To further examine the nature of any observed set shifting deficit, we next applied a computational model to the data of the AN-WR women and control participants to determine if the pattern of results could be simulated by altering key parameters in the model that represent behavioral and neural processes impacted in AN. Many alternative computational models exist that could potentially be used for this purpose. However, during the past decade, two developments have significantly narrowed the set of viable models. First, there are now many results suggesting that human categorization is mediated by multiple category-learning systems (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & O'Brien, 2005; Erickson & Kruschke, 1998; Love, Medin, & Gureckis, 2004; Reber, Gitelman, Parrish, & Mesulam, 2003). Much of this evidence is in the form of behavioral dissociations between rule-based and information-integration category-learning tasks. In rule-based category-learning tasks, the categories can be separated by a rule that is easy to describe verbally and that often only depends on a single stimulus dimension. One widely known example is the WCST. In contrast, accurate performance in an information-integration task requires an implicit integration of two or more stimulus dimensions in a way that is non-verbalizable. At least 20 qualitatively different behavioral dissociations have been reported (with healthy young adults) between rule-based and information-integration categorization (e.g., Ashby & Maddox, 2005, 2010). These results have profoundly affected the categorization field, partly because no single-system theory has been able to account for more than one or two of these dissociations simultaneously. Second, there has been an explosion of new knowledge about the neural basis of category learning (Ashby & Ennis, 2006; Ashby, Noble, Filoteo, Waldron, & Ell, 2003; Filoteo & Maddox, 2007; Maddox & Filoteo, 2005, 2007; Nomura et al., 2007; Nomura & Reber, 2008; Seger, 2008; Seger & Cincotta, 2005, 2006). These new data come from a variety of sources, including fMRI, EEG, single-unit recordings, and behavioral studies with a variety of different neuropsychological patient populations. Purely cognitive models make no predictions about any of these new data. In fact, to date, the only theory of category learning that makes central the constraints imposed by the underlying neurobiology is the COVIS model (Ashby et al., 1998, Ashby, Paul, & Maddox, 2011). At the same time, COVIS is the only model that can account for all of the rule-based versus information-integration dissociations that have been reported.

COVIS assumes that there are two neurocognitive systems that compete throughout the learning of novel categories – an explicit, hypothesis-testing system and an implicit, procedural-learning system. The explicit system uses logical reasoning and depends on working memory and executive attention and is mediated by a broad neural network that includes prefrontal cortex, anterior cingulate, hippocampus, and frontal-striatal circuits comprised by the head of the caudate. The implicit system depends on posterior striatal regions and areas of premotor cortex and learns via dopamine-mediated synaptic plasticity at cortical-striatal synapses.

The computational version of COVIS includes three main components – one that models the hypothesis-testing system, one that models the procedural-learning system, and a third that monitors the output of the two systems and selects a categorization response. For a complete description of the model, see Ashby et al. (2011). Briefly, however, the hypothesis-testing component selects and tests explicit rules that determine category membership (e.g., one-dimensional rules). Computationally, this component of COVIS is implemented as a hybrid neural network that includes both symbolic and connectionist features. On each trial, the most salient of all possible rules is selected for application. Rule salience, which is updated every time feedback is provided, is a function of life history (initial bias), categorization accuracy (i.e., the proportion of times that positive feedback follows rule use), a tendency to perseverate, and a tendency to experiment with novel rules. The procedural-learning component is implemented as a three-layer feed forward connectionist network with up to 10,000 units in the sensory (i.e., input) layer, one unit in the hidden layer for each contrasting category, and one unit in the motor (i.e., output) layer for each response alternative. Learning occurs at synapses between the input and hidden layer units and follows reinforcement learning rules. COVIS has recently been successfully applied to simulate the

pattern of cognitive deficits observed in patients with Parkinson's disease, who are well known to have lowered dopamine levels and also present with set shifting deficits (Hélie, Paul, & Ashby, 2012a, 2012b).

The present study used a rule-based category learning task, in which the COVIS hypothesis-testing system is predicted to dominate. Therefore, this component of the model is most critical in the current application. The Appendix describes the COVIS hypothesis-testing system in detail, but briefly, this component of COVIS has several important parameters, but those most relevant to AN are the $\Delta_e$parameter, which measures sensitivity to negative feedback, the parameter, which measures the tendency to select unusual rules, and the parameter, which measures the tendency to perseverate with the same rule, even in the face of negative feedback. These parameters are important in the present application of COVIS to category learning and set shifting in AN for the following reasons. Sensitivity to negative feedback or punishment (instantiated by the $\Delta_e$parameter) is relevant given past studies demonstrating that AN patients have altered reward sensitivity (Davis & Woodside, 2002; Keating, 2010), in general, and increased sensitivity to punishment (Harrison, O'Brien, Lopez, & Treasure, 2010; Jappe et al., 2011), in particular. In fact, there appears to be an abnormal response to feedback in adult and adolescent AN patients based on fMRI (Bischoff-Grethe, Hazeltine, Bergren, Ivry, & Grafton, 2009; Fladung et al., 2010; Wagner et al., 2007).

Although AN is not specifically associated with systemic dopamine loss, a number of neurotransmitter systems including dopamine may be disrupted in the disorder (Avena & Bocarsly, 2012). For example, it has been observed that recovered AN patients show aberrant DA activity (Kaye, Frank, & McConaha, 1999) and increased D2/D3 dopamine receptor binding (Frank et al., 2005). Wagner and colleagues (2007) further observed that the fMRI BOLD response in ventral striatum (heavily interconnected to the midbrain dopaminergic system) fails to differentiate between positive and negative feedback in recovered AN patients, but clearly differentiates between these different types of feedback in normal controls. This further suggests general involvement of reward processing and dopamine-related mechanisms in the disease; however, experimental evidence is contradictory as to whether these aberrations are characterized by increases or decreases in the dopamine system (for a review, see Kontis & Theochari, 2012). Modeling AN patient data using COVIS could help resolve this controversy and thereby refine AN treatment strategies.

Considerable evidence suggests that creative problem solving improves with brain dopamine levels and the tendency to perseverate is reduced (Ashby, Isen, & Turken, 1999). COVIS accounts for these effects via separate rule selection and rule switching parameters that change in opposite ways as dopamine levels rise. Specifically, in COVIS, as dopamine levels rise in the anterior cingulate and prefrontal cortex, the parameter increases, which increases the probability that a low-salience rule is selected on any given trial (and via this mechanism, creative problem solving is improved). In contrast, the tendency to ignore feedback and perseverate on the current rule increases with the parameter, which is inversely related to levels of striatal dopamine. Given that set shifting has been associated with brain dopamine function (Floresco, Ghods-Sharifi, Vexelman, & Magyar, 2006; Kehagia, Barker, & Robbins, 2010) and that the pathophysiology of AN may involve dopamine alterations (e.g., Frank et al., 2005; Kaye, et al.,, 1999; and as discussed above), it is possible that alterations in set-shifting in AN represent changes to the dopamine system. In regard to this latter point, and as noted earlier, COVIS has been successful in simulating the set shifting deficits observed in patients with Parkinson's disease who have noted depletions in dopamine levels (Hélie et al., 2012a, 2012b).

Based on previous work, we predicted that the weight-restored AN participants would demonstrate impaired set shifting abilities in that once the rule shifts, they would continue to perseverate on the previous rule. Further, based on the notion that AN patients experience alterations in responding to negative feedback and dopamine systems, we anticipated that systematic manipulations of specific parameters of COVIS would provide a good accounting of the AN's performance on the set-shifting task.

Method

Participants

Nineteen AN-WR women and 35 control women (CW) participated in this study. Participants in the AN-WR group were recruited through the University of California San Diego Eating Disorders program. Participants with AN-WR previously met DSM-IV-TR (APA, 2000) criteria for anorexia nervosa; 12 with restricting subtype, 5 with purging subtype, and 2 with binging subtype. To be considered weight-restored, an AN-WR participants had to have maintained 90% of their ideal body weight for at least one year. AN-WR participants also had to have regular menstrual cycles, not use psychoactive medication, or engage in abnormal eating behaviors (e.g., restricting patterns) for at least one year. CW participants were recruited through local advertisements in the San Diego (n=19) and Denver (n=16) metropolitan areas. The two CW groups did not differ in demographics. CW participants had a lifetime history of body weight between 90% and 110% of ideal body weight since menarche, and had no history of psychiatric or major medical illness. A doctoral level clinician assessed AN-WR and CW

participants with the Structured Clinical Interview for DSM-IV Axis I Disorders (First, Gibbon, Spitzer, & Williams, 1996). Five AN-WR participants had comorbid Major Depressive Disorder and one had a comorbid anxiety disorder; no individuals had a psychotic, substance use or bipolar disorder. The results reported below were reanalyzed after first removing those AN-WR participants with comorbid depression and then removing the AN-WR participant with comorbid anxiety and the pattern of the results did not differ as compared to when the entire sample was included. Study participants completed the Temperament Character Inventory (Cloninger, Przybeck, Svrakic, & Wetzel, 1994) and the Eating Disorder Inventory – 2 (EDI-2; Garner, 1991) at the time of this study.

Table 1 displays the means and standard deviations of participants age at the time of study, age at illness onset, age when weight was restored to normal, duration of low weight, duration weight had been restored to normal, current body mass index (BMI), lowest BMI, difference between lowest and current BMI, and scores on selected subscales of the TCI and EDI-2 for the all participants in the AN-WR and CW groups (when applicable). Written informed consent was obtained for each participant after a complete description of the study procedures was provided. The local Institutional Review Boards approved all research procedures.

Stimuli

The category learning task was adapted from that used by Filoteo and colleagues to study explicit category learning and set-shifting in patients with basal ganglia disorders (Filoteo, Maddox, Ing, Zizak, & Song, 2005). Two different sets of computer-generated stimuli were presented consisting of color images of either cartoon "castles" or "houses". Examples of stimuli from each of the two sets are shown in Figure 1. For each set, four possible binary-valued dimensions could vary from trial-to-trial. These four dimensions and the binary values for each stimulus set were the following: *castle stimuli* - shape of foundation (diamond or square), location of ramparts (above walls or sunken into walls), number of rings surrounding castle (1 or 2), color of drawbridge (yellow or green); *house stimuli* - color of door (blue or red), lighting inside window (lights off or lights on), shape of roof (flat or triangular), nature of plants (shrub or tree). Each stimulus was presented in color that remained constant except for the altered dimension that was relevant to the categorization task described above. Each stimulus was approximately 10 cm in height and from a viewing distance of approximately 60 cm subtended about 9.6 degrees of visual angle.

Procedure

Each participant was randomly administered one set of stimuli (houses or castles). Participants were told that they would be shown individual pictures and asked to categorize each as either belonging to Category 1 or Category 2 by pressing a specified key. Participants were also told that after they categorized the picture, they would receive feedback on the computer screen in the form of the word “Correct” for correct responses and the word “Wrong” for incorrect responses. Participants were also told that they would be guessing at first and that they should attempt to learn from their errors. For each set of stimuli, four dimensions would vary on a trial-by-trial basis. The task of the participant was to determine the relevant dimension based on the corrective feedback. Participants were presented a total of 160 trials in 8 blocks: 80 pre rule-shift trials and 80 post rule-shift trials. For the castle stimuli, the relevant dimension prior to the rule-shift was the shape of the foundation (Category 1 = square shape, Category 2 = diamond shape), and the relevant dimension after the rule-shift was the number of rings around the castle (Category 1 = one ring, Category 2 = two rings). For the house stimuli, the relevant dimension prior to the rule-shift was the shape of the roof (Category 1 = flat, Category 2 = triangular), and the relevant dimension after the rule-shift was the nature of the plants (Category 1 = tree, Category 2 = bush). Participants were never informed that a rule-shift was going to occur and thus had to infer it based on the corrective feedback.

Each trial began with the presentation of a picture that remained on the screen until the participant made a categorization response. Immediately following a response, correct or incorrect feedback was presented for 0.75 sec while the stimulus remained on the screen, followed by a blank screen for 1.0 sec, and then the presentation of the next stimulus.

Statistical Analysis of Behavioral Data

Demographic and clinical variables were compared using independent sample t-tests. Accuracy performances (proportion correct) on the category learning task were examined using group × block, mixed-design ANOVAs. To determine initial rule acquisition, a Learning Slope index was computed for each participant by subtracting the proportion correct on block 1 from the proportion correct on block 4 (greater slopes equaled

greater rule acquisition). To determine the impact of the rule shift, a Shift Cost score was computed by subtracting each participant's proportion correct on block 5 from their proportion correct on block 4 (greater scores equaled a greater shift-cost). The Learning Slope and Shift Cost scores for the patient and control groups were then compared using independent sample t-tests. Correlation analyses were conducted using Pearson's correlation analysis. All statistical tests were two-tailed and considered reliable at the $p < .05$ level. We did not correct for multiple ANOVA tests because each analyses addressed separate *a priori* questions and ANOVA generally tends to protect from Type I error. Effect sizes are reported as either Cohen's *d* or partial eta squared η_p^2.

COVIS Simulations

The computational implementation of COVIS employed here is described fully by Ashby et al. (2011). To simulate the observed behavioral performance, COVIS parameters were adjusted systematically between groups and with respect to the theoretically motivated contribution of each parameter. As explained earlier, three parameters potentially relevant to AN patients were manipulated to account for differences between AN and CW participants: the perseveration parameter γ (hypothesized to be inversely related to striatal dopamine levels), the selection parameter λ (directly related to cortical dopamine levels), and Δ_e, which measures sensitivity to error feedback. In the following simulations, γ was larger for AN simulations than control simulations, while λ was larger for control than AN simulations. This relationship reflects the hypothesized role of dopamine deficiency in AN with respect to COVIS. Finally, Δ_e was larger for AN than control simulations, in line with the observation that AN patients are more sensitive to punishment.

The parameters were constrained to an ordinal relationship across participant populations with respect to the presumed dopamine-related effects of AN. All parameters in the model were initially estimated by fitting the model to data from the CW group, and only the parameters described above were modified to fit data from the AN-WR group. None of the parameter estimates were optimized; reasonable values were assigned using a rough grid search.

Two hundred simulations were run for each control and AN-WR participant in the set-shifting task described above. As in prior applications of COVIS, the procedural system received an object-based representation of the stimuli as a 16-dimensional binary vector: for stimulus *i*, the vector had a value of 1 in position *i* and a value of 0 otherwise. The hypothesis-testing system received a feature-based representation of the stimuli as a four-dimensional binary vector: for each stimulus, the entry in row *j* corresponding to feature *j* (e.g., door) was set to 1 if it had one value (e.g., blue) and 0 if it had the other (e.g., red).

Results

Demographic, TCI-2, and EDI-2 Comparisons

Demographic and disease characteristics are displayed in Table 1. The mean age of the AN-WR and CW groups did not differ nor did their BMI. On the TCI-2, the AN group reported greater scores on the Harm Avoidance and Persistence subscales than the CW group indicating that the AN-WR group was more harm avoidant and persistent in their personalities. The two groups differed significantly on all of the EDI-2 subscales examined indicating that, compared to the CW group, the AN group had greater drive for thinness, bulimia symptoms, body dissatisfaction, feelings of ineffectiveness, and perfectionistic tendencies.

Behavioral Accuracy Results

Accuracy results (proportion correct) across the 160 trials in 8 trial blocks are depicted in Figure 2. The groups were contrasted separately on blocks 1-4 (pre-shift blocks) and blocks 5-8 (post-shift blocks) using two separate 2 (group) × 4 (block) ANOVAs to determine if there were any differences between the groups in rule learning (blocks 1-4) and set shifting (blocks 5-8), respectively. The results for blocks 1-4 revealed a group x block interaction, $F(3,156)=5.49$, $p=0.001$, $\eta_p^2=0.096$, such that the AN group demonstrated greater learning across the four blocks than the CW group (see Figure 2). The ANOVA also revealed an effect of block, $F(3,156)=11.13$, $p<0.001$, $\eta_p^2=0.176$, but no effect of group, $F(1,52)=0.69$, $p=0.41$, $\eta_p^2=0.013$. The significant group X block interaction observed in blocks 1-4 was further supported by an examination of the two groups' Learning Slopes. The mean Learning Slope for the AN-WR group was 31.8 (SD=15.4) and for the CW group it was 16.1 (SD=22.3), and these were significantly different, $t(52)=2.73$, $p=0.009$, $\eta_p^2=0.125$.

The results for the 2 X 4 ANOVA for blocks 5-8 revealed an effect of group, $F(1,52)=4.66$, $p=0.036$, $\eta_p^2=0.082$, and block, $F(3,156)=9.98$, $p<0.001$, $\eta_p^2=0.161$, but no group × block interaction, $F(3,156)=0.51$, $p=0.68$, $\eta_p^2=0.010$. To examine the cost in accuracy following the rule shift, a 2 (group) × 2 (block) ANOVA was used to contrast the groups' accuracies in block 4 vs. block 5. The results for this analysis revealed a significant group × block interaction, $F(1,52)=4.17$, $p=0.046$, $\eta_p^2=0.074$, and a significant effect of block, $F(1,52)=55.32$, $p<0.001$, $\eta_p^2=0.515$, but no effect of group, $F(1,52)=0.02$, $p=0.89$, $\eta_p^2=0.000$. We next contrasted the mean Shift-Cost for the AN group (32.6, *SD*=22.3) and for the CW group (18.6, *SD*=25.0) and these means were significantly different, $t(52)=2.04$, $p=0.046$, $\eta_p^2=0.074$.

An examination of Figure 2 suggests that one possible reason the AN group displayed a greater Shift-Cost than the CW group was that they learned the rule prior to the shift to a greater extent, thereby making it more difficult to shift once the rule changed. That is, it could have been that the AN-WR participants became "locked in set" prior to the rule change by virtue of their initial hyper learning. To test this hypothesis, we correlated participants' Learning Slope and Shift Cost and found that these did not correlate in the AN-WR group ($r(19)=0.10$, $p=0.67$) but there was a trend for an association in the CW group ($r(35)=0.33$, $p=0.057$).

Clinical Correlates

We next examined clinical correlations between the task variables and the clinical variables in the AN-WR and CW groups. These correlations are shown in Table 2. For the AN-WR group, a lower duration of weight restoration and a lower change in BMI was associated with a greater, more abnormal Learning Slope value (i.e., hyper-learning), and greater scores on the Harm Avoidance subscale of the TCI was associated with a greater, more abnormal Learning Slope. There were no significant associations between any of the demographic or clinical variables and Shift Cost for the AN-WR group. For the CW group, older age was associated with a greater Learning Slope (faster learning), but no other variables, and Shift Cost was not associated with any of these variables.

Simulation Results

The exact parameter values for each participant group appear in Tables 3 and 4. Note that only the three parameters described earlier were adjusted to simulate the behavioral performance of the AN-WR and CW participants—the remaining parameters required for the model were set to identical values across both groups, and are described elsewhere (Ashby et al., 2011; Hélie et al., 2012a, 2012b).

Figure 3 shows the simulated performance of the model in the behavioral task. Note that the absence of error bars is for two reasons: first, the error bars can be made arbitrarily small by increasing the number of simulations; second, COVIS does not specifically model individual variability in task performance. For the simulated adults, AN-CW patients learn faster in the first four blocks due to the higher sensitivity to negative feedback and the model also captures the greater switch cost (block 4 – block 5 accuracy) for the AN group. Post-switch, the simulated control group ends at a higher accuracy than the AN group due to the perseverative tendency of the AN group. Overall, COVIS closely matches the human performance data suggesting dopamine-decreases may, at least in part, underlie the cognitive deficits observed in this task.

To evaluate the robustness of the simulated performance to changes in the parameters, a sensitivity analysis (Hélie et al., 2012a, 2012b) was carried out. Briefly, the value of each parameter used to fit the behavioral data was adjusted ±10% and then ±95% in turn for an additional 200 simulations for each parameter change. Next, for each change, new predictions were generated and the average difference between the Figure 3 predictions and the new predictions was quantified by calculating the mean root squared error (MRSE). Across all simulated groups and every parameter, this analysis yielded an overall MRSE of only 2.46% when the parameters were changed by ±10%, and an MRSE of 5.66% when the parameters were changed by ±95%. These small MRSEs show that the model tends to make highly similar predictions even when the parameters change substantially. For this reason, we can be confident that an optimized parameter estimation process would not significantly alter the model's predictions, relative to the course grid search that we used.

COVIS is insensitive to small changes in parameter values because the model is richer in structure than more typical purely cognitive models. For example, in the present application COVIS assumes that participants always experiment with only four simple explicit rules – namely, the one-dimensional rules on each of the four stimulus dimensions. Thus, no matter how the parameters are changed the model always predicts that one of these four strategies must be applied on every trial. Changing the parameters simply changes the probabilities that each of the four rules is selected. This can change the predictions of the model quantitatively, but it cannot change them qualitatively. In contrast, consider a purely cognitive model of categorization such as the exemplar-based,

generalized context model (GCM; Nosofsky, 1986). By manipulating its (attention weight) parameters, the GCM can mimic any of the four rule strategies assumed by COVIS, but it can also mimic many more complex decision strategies that integrate information from multiple dimensions. Thus, in contrast to COVIS, changing the GCM parameters can change the qualitative predictions of the model as well as the quantitative predictions.

This same sensitivity analysis also allowed us to ascertain that the the rule selection parameter (i.e., λ; hypothesized to be directly related to cortical DA levels) accounts for the most variability in the model's performance. For example, the variance of the MRSEs that result when each parameter is changed by ±95% was 40.8 for the selection parameter, 9.0 for the perseveration parameter (i.e., γ), and 1.3 for the sensitivity-to-error-feedback parameter (i.e., Δ_e). This result is important because it allows us to pinpoint what micro-process accounts for the performance difference between restored AN and CW. Specifically, the COVIS simulations suggest that recovered AN do not necessarily have a deficit in disengaging from an unsuccessful response strategy, but instead the deficit may be caused by a difficulty in selecting a new strategy to replace the old (unsuccessful) strategy. This difficulty in selecting a new strategy may be related to AN's perfectionist personality (Casper et al., 1992; Merwin et al, 2011), as AN may be worried about selecting an even worse strategy, and receiving more negative feedback (i.e., the next strategy needs to be the correct one). Future work is needed to specifically test for this new hypothesis.

Discussion

A main finding from the present study is that, compared to CW participants, AN-WR participants were impaired in set shifting. These results are consistent with previous studies demonstrating a set shifting deficit in patients both currently ill with an eating disorder (Roberts et al., 2007; Roberts et al., 2010; Shott et al., 2012; Steinglass et al., 2006; Tchanturia et al., 2002) and following weight restoration (Danner et al., 2012; Roberts et al., 2010; Tchanturia et al., 2002; Tenconi et al., 2010).

A novel behavioral finding in the current study was that the AN-WR group demonstrated hyper learning during the initial rule acquisition stage prior to the rule shift. This finding is consistent with previous studies showing that AN patients are better at focusing on the detailed aspects of visual information (Lopez, Tchanturia, Stahl, & Treasure, 2009; Southgate, Tchanturia, & Treasure, 2008) which is an important aspect of rule-based learning (Ashby et al., 1998). That is, during the first stage of the category learning task the participant had to identify one stimulus dimension and ignore the other, irrelevant dimensions to acquire the rule. If AN-WR participants are better able to focus on details of visual objects at the cost of the overall gestalt (often described as a deficit in central coherence) they could likely learn the rule better than CW participants. However, when the rule changes, AN-WR groups' deficit in rule shifting manifests and they are unable to learn the new rule. This explanation of our findings proposes that two separate mechanisms are responsible for the hyper learning and set shifting deficits observed in the present study, which is highly consistent with our finding of a lack of a correlation between AN-WR participants' Learning Slope and Set Shift Cost values.

The results from the model simulations are also consistent with the notion that two distinct mechanisms underlie AN-WR participants' hyper learning and set shifting deficits. Specifically, increasing the value of the parameter Δ_e, which measures sensitivity to error feedback, resulted in hyper learning during the initial stages of rule acquisition but had no impact on the ability of the model to simulate AN-WR participants' set shifting deficit. In contrast, increasing the perseveration parameter γ, which is hypothesized to be inversely related to striatal dopamine levels, and decreasing the selection parameter λ, which is thought to be directly related to cortical dopamine levels, resulted in an increase in perseverations that simulated the behavioral set shifting deficit observed in the AN group (see Figure 3). As such, COVIS provided a good accounting of all aspects of the results observed in the AN group and the model suggests that different mechanisms underlie the two behavioral findings. Of note, the parameters that were manipulated were not selected arbitrarily and were based on our understanding of both the behavioral and neurobiological findings in AN. That is, AN patients are more sensitive to negative feedback than control participants (Harrison et al., 2010; Harrison, Treasure, & Smillie, 2011; Jappe et al., 2011) and this sensitivity to punishment has been linked with abnormal eating behaviors in non-clinical samples (Loxton & Dawe, 2006). As for those parameters thought to be sensitive to the integrity of dopamine levels (γ and λ), these were selected given previous studies showing increased dopamine binding using PET imaging (indicative of decreased dopamine levels) in the ventral striatum in weight restored AN patients (Bailer et al., 2013; Frank et al., 2005). Taken together, the simulation results are very promising in terms of helping to better pinpoint the nature of set shifting performance in AN-WR individuals and to help support the hypothesis that alterations in striatal dopamine may be the underlying neurobiological substrate for set shifting deficits in these patients. Of course, more direct neurobiological data will be needed to help further support this possibility.

We also examined the clinical correlates of AN patients' hyper learning and set shifting deficit by correlating participants' Learning Slope and Set Shift Cost values with their clinical variables and scores on selected TCI-2 and ED-2 subscales. Interestingly the results of these correlations also suggest that hyper learning and set

shifting deficits in the AN-WR group are due to distinct mechanisms. Specifically, a shorter duration of weight restoration was associated with a greater, more abnormal learning slope, as was a smaller change in BMI between lowest and current BMI. These findings indicate that recovery status may be associated with hyper learning in AN-WR individuals. However, it was also the case that greater scores on the Harm Avoidance subscale of the TCI-2 was associated with a larger, more abnormal learning slope. One of the features of individuals who obtain high scores on the Harm Avoidant subscale is that they tend to be highly sensitive to criticism and punishment, suggesting that this personality feature might underlie the hyper learning observed in our sample of AN-WR participants. This is highly consistent with the results of the modeling analyses and further indicates the utility of our modeling approach in providing converging evidence as to the mechanisms underling our behavioral findings.

In contrast to the findings with the Learning Slope, the Shift Cost index was not associated with any of the clinical variables or subscales from the TCI-2 or EDI-2. This is consistent with previous studies that also did not find significant associations in currently ill AN patients between set shifting and BMI and disease duration (Shott et al., 2012; Tchanturia et al., 2011) or between set shifting and subscales of the TCI-2 or ED-2 (Shott et al., 2012). These results are important and raise two important issues in AN research. First, given that the various BMI and disease duration variables are not associated with set shifting deficits in the current study, it appears that disease severity does not necessarily account for the set shifting deficit in our AN-WR sample. If so, then our findings may not be attributable to the neurological impact of the acute stages of AN or possibly any long-lasting effects of the disease but may reflect cognitive traits of AN seen in adulthood.

The other important issue raised by the lack of correlations among the Shift Cost indices and the clinical measures is whether there is any clinical utility to identifying set shifting deficits in AN-WR individuals. That is, if the personality characteristics that are thought to be the hallmark of AN are not associated with set shifting in AN, how meaningful is this cognitive alteration. For example, AN patients are often rigid, perfectionistic, and harm avoidant (Bastiani, Rao, Weltzin, & Kaye, 1995; Friederich & Herzog, 2011) and these characteristics can persist after weight restoration (Klump et al., 2004) and can predict important treatment variables such as response and drop out (Fassino et al., 2005; Pham-Scottez et al., 2012). The lack of an association between set shifting deficits and these characteristics (as measured by the TCI) in weight restored AN participants and currently ill AN patients (Shott et al., 2012) highlights this concern. Thus, while the finding of a set shifting deficit in AN is interesting from a cognitive neuropsychiatric perspective, the clinical utility of this cognitive characteristic awaits further research. More work is clearly needed to determine if hyper learning and/or set shifting deficits are predictive of important disease and treatment variables such as disease severity, treatment participation, treatment outcome, and relapse, to name a few. However, the finding that AN-WR participants demonstrate both hyper learning and set shifting deficits, that these two abnormal behaviors are not associated, that the clinical correlates of these behaviors differ, and that these two behaviors are simulated with different model parameters in COVIS, provides great encouragement as to the possible clinical utility of identifying either hyper learning or set shifting deficits in individuals with AN.

There are two important limitations to the study that should be addressed. First, the small sample size is one limitation making it difficult to generalize our findings to all weight-restored AN individulas. However, the observation of a set shifting deficit in AN (both currently ill and weight restored) is one of the most consistent findings in eating disorders research, which strengthens our contention that our current findings are reliable and generalizable. In contrast, the observation of hyper learning in our AN sample is a novel finding that will require replication in larger samples. Second, the design of the study was cross sectional making it difficult to determine the evolution of hyper learning and set shifting deficits in AN and the contribution of these cognitive characteristics to developing the disease. That is, based on the present study we cannot determine whether hyper learning or the set shifting deficits contribute to the development of AN or are a result of the disease. Of note, in a recent study we found that currently ill adolescents with AN did not demonstrate hyper learning or set shifting deficits on the same task as the one used in the current study (Shott et al., 2012), suggesting that set shifting deficits may not necessarily contribute to the onset of AN. However, it is still not clear whether set shifting deficits are entirely normal in adolescents with AN so it currently cannot be concluded that hyper learning or set shifting deficits do not contribute to the development of AN.

In summary, the current study identified a set shifting deficit in weight restored AN patients, a finding highly consistent with previous work. In addition, AN-WR participants demonstrated hyper learning compared to CW participants, which has not been reported previously. Both of these findings were accurately simulated by COVIS, a biologically plausible model of category learning and set shifting, by manipulating model parameters that represent sensitivity to punishment and dopamine functioning, neural processes known to be impacted in AN. The clinical utility of these findings awaits further study but the results of this study provide great promise for the use of computational modeling in better understanding neuropsychological functioning in AN.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-81.

Ashby, F.G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *The Psychology of Learning and Motivation*, *46*, 1-36.

Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*, 529-550.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149-178.

Ashby, F. G., & Maddox, W. T. (2010). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224,* 147-161.

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., & Ell, S. W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology, 17*, 115-24.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *TRENDS in Cognitive Science, 2*, 83-89.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A.J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 65-87). New York: Cambridge University Press.

Avena, N. M., & Bocarsly, M. E. (2012). Dysregulation of brain reward systems in eating disorders: Neurochemical information from animal models of binge eating, bulimia nervosa, and anorexia nervosa. *Neuropharmacology*, *63*, 87–96. doi:10.1016/j.neuropharm.2011.11.010

Bailer, U. F., Frank, G. K., Price, J. C., Meltzer, C. C., Becker, C., Mathis, C. A., et al. (2013). Interaction between serotonin transporter and dopamine D2/D3 receptor radioligand measures is associated with harm avoidant symptoms in anorexia and bulimia nervosa. *Psychiatry Research, 211*, 160-168.

Bastiani, A. M., Rao, R., Weltzin, T., & Kaye, W. H. (1995). Perfectionism in anorexia nervosa. *International Journal of Eating Disorders, 17*, 147-152.

Bischoff-Grethe, A., Hazeltine, E., Bergren, L., Ivry, R. B., & Grafton, S. T. (2009). The influence of feedback valence in associative learning. *Neuroimage, 44*, 243-51.

Casper, R. C., Hedeker, D., & McClough, J. (1992). Personality dimensions in eating disorders and their relevance for subtyping. *Journal of the American Academy of Child & Adolescent Psychiatry, 31*, 830-840.

Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., & Wetzel, R. D. (1994). *The temperament and character inventory (TCI): A guide to its development and use* Center for Psychobiology of Personality, Washington University St. Louis, MO.

Danner, U.N., Sanders, N., Smeets, P.A.M., van Meer, F., Adan, R.A.H., Hoek, H.W., & van Elburg, A.A. (2012). Neuropsychological weaknesses in anorexia nervosa: Set-shifting, central coherence and decision making in currently ill and recovered women. *International Journal of Eating Disorders*, 45, 685-694.

Davis, C., & Woodside, D. B. (2002). Sensitivity to the rewarding effects of food and exercise in the eating disorders. *Comprehensive Psychiatry, 43*, 189-194.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Fassino, S., Abbate Daga, G., Delsedime, N., Busso, F., Piero, A., & Rovera, G. G. (2005). Baseline personality characteristics of responders to 6-month psychotherapy in eating disorders: Preliminary data. *Eating and Weight Disorders : EWD, 10*, 40-50.

Filoteo, J.V. & W.T. Maddox. (2007). Category learning in Parkinson's disease. In Research progress in Alzheimer's disease and dementia. Maio-Kun Sun, Ed.: Vol. 3, 2–26. Nova Science Publishers, Inc. Hauppauge, NY.

Filoteo, J. V., Maddox, W. T., Ing, A. D., Zizak, V., & Song, D. D. (2005). The impact of irrelevant dimensional variation on rule-based category learning in patients with Parkinson's disease. *Journal of the International Neuropsychological Society, 11*, 503-13.

First, M. B., Gibbon, M., Spitzer, R., & Williams, J. (1996). User's guide for the structured clinical interview for DSM-IV axis I Disorders—Research version. *SCID-I, Version, 2.*

Fladung, A. K., Grön, G., Grammer, K., Herrnberger, B., Schilly, E., Grasteit, S., et al. (2010). A neural signature of anorexia nervosa in the ventral striatal reward system. *American Journal of Psychiatry, 167*, 206-212.

Floresco, S. B., Ghods-Sharifi, S., Vexelman, C., & Magyar, O. (2006). Dissociable roles for the nucleus accumbens core and shell in regulating set shifting. *The Journal of Neuroscience, 26*, 2449-2457.

Frank, G. K., Bailer, U. F., Henry, S. E., Drevets, W., Meltzer, C. C., Price, J. C., et al. (2005). Increased dopamine D2/D3 receptor binding after recovery from anorexia nervosa measured by positron emission tomography and [< sup> 11</sup> C] raclopride. *Biological Psychiatry, 58*, 908-912.

Friederich, H. C., & Herzog, W. (2011). Cognitive-behavioral flexibility in anorexia nervosa. *Behavioral Neurobiology of Eating Disorders, 6,* 111-123.

Garner, D. M. (1991). *Eating disorder inventory-2: Professional manual* Psychological Assessment Resources Odessa, FL.

Harrison, A., O'Brien, N., Lopez, C., & Treasure, J. (2010). Sensitivity to reward and punishment in eating disorders. *Psychiatry Research, 177,* 1-11.

Harrison, A., Treasure, J., & Smillie, L. D. (2011). Approach and avoidance motivation in eating disorders. *Psychiatry Research, 188*, 396-401.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012a). Simulating the effects of dopamine imbalance on cognition: From positive affect to Parkinson's disease. *Neural Networks*, *32*, 74–85.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012b). A neurocomputational account of cognitive deficits in Parkinson's disease. *Neuropsychologia, 50*, 2290-2302.

Jappe, L. M., Frank, G. K. W., Shott, M. E., Rollin, M. D. H., Pryor, T., Hagman, J. O., et al. (2011). Heightened sensitivity to reward and punishment in anorexia nervosa. *International Journal of Eating Disorders, 44*, 317-324.

Kaye, W. H., Frank, G. K. W., & McConaha, C. (1999). Altered dopamine activity after recovery from restricting-type anorexia nervosa. *Neuropsychopharmacology, 21*, 503-506.

Keating, C. (2010). Theoretical perspective on anorexia nervosa: The conflict of reward. *Neuroscience & Biobehavioral Reviews, 34*, 73-79.

Kehagia, A. A., Barker, R. A., & Robbins, T. W. (2010). Neuropsychological and clinical heterogeneity of cognitive impairment and dementia in patients with Parkinson's disease. *Lancet Neurology, 9*, 1200-1213.

Klump, K. L., Strober, M., Bulik, C. M., Thornton, L., Johnson, C., Devlin, B., et al. (2004). Personality characteristics of women before and after recovery from an eating disorder. *Psychological Medicine, 34*, 1407-1418.

Kontis, D., & Theochari, E. (2012). Dopamine in anorexia nervosa: a systematic review. *Behavioural Pharmacology*, *23*, 496–515. doi:10.1097/FBP.0b013e328357e115

Lopez, C., Tchanturia, K., Stahl, D., & Treasure, J. (2009). Weak central coherence in eating disorders: A step towards looking for an endophenotype of eating disorders. *Journal of Clinical and Experimental Neuropsychology, 31*, 117-125.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*, 309-332.

Loxton, N. J., & Dawe, S. (2006). Reward and punishment sensitivity in dysfunctional eating and hazardous drinking women: Associations with family risk. *Appetite, 47*, 361-371.

Maddox, W. T., & Filoteo, J. V. (2005). The neuropsychology of perceptual category learning. In H. Cohen & C. Lefebvre (Eds.), Handbook of categorization in cognitive science (pp. 573-599). Elsevier, Ltd.

Maddox, W. T., & Filoteo, J. V. (2007). Modeling visual attention and category learning in amnesiacs, striatal-damaged patients and normal aging. In R. W. J. Neufeld (Ed.), Advances in clinical cognitive science: Formal modeling and assessment of processes and symptoms (pp. 113-146). Washington DC: American Psychological Association.

Merwin, R. M. (2011). Anorexia nervosa as a disorder of emotion regulation: Theory, evidence, and treatment implications. *Clinical Psychology: Science and Practice, 18*, 208-214.

Merwin, R. M., Timko, C. A., Moskovich, A. A., Ingle, K. K., Bulik, C. M., & Zucker, N. L. (2011). Psychological inflexibility and symptom expression in anorexia nervosa. *Eating Disorders*, 19, 62-82.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, A. D., Gitelman, D. R., Parrish, T. B., et al. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17*, 37-43.

Nomura, E. M., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience and Biobehavioral Reviews*, *32*, 279-291.

Pham-Scottez, A., Huas, C., Perez-Diaz, F., Nordon, C., Divac, S., Dardennes, R., et al. (2012). Why do people with eating disorders drop out from inpatient treatment?: The role of personality factors. *The Journal of Nervous and Mental Disease, 200*, 807-813.

Reber, P. J., Gitelman, D. R., Parrish, T. B., & Mesulam, M. M. (2003). Dissociating explicit and implicit category knowledge with fMRI. *Journal of Cognitive Neuroscience*, *15*, 574-583.

Roberts, M. E., Tchanturia, K., Stahl, D., Southgate, L., & Treasure, J. (2007). A systematic review and meta-analysis of set-shifting ability in eating disorders. *PSYCHOLOGICAL MEDICINE-LONDON-, 37*, 1075.

Roberts, M. E., Tchanturia, K., & Treasure, J. L. (2010). Exploring the neurocognitive signature of poor set-shifting in anorexia and bulimia nervosa. *Journal of Psychiatric Research, 44*, 964-970.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Review*s, *32*, 265-278.

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *Journal of Neuroscience*, *25*, 2941-2951.

Seger, C. A., & Cincotta, C. M. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cerebral Cortex*, *16*, 1546-1555.

Shott, M. E., Filoteo, J. V., Bhatnagar, K. A., Peak, N. J., Hagman, J. O., Rockwell, R., et al. (2012). Cognitive set-shifting in anorexia nervosa. *European Eating Disorders Review: The Journal of the Eating Disorders Association, 20*, 343-349.

Southgate, L., Tchanturia, K., & Treasure, J. (2008). Information processing bias in anorexia nervosa. *Psychiatry Research, 160*, 221-7.

Steinglass, J. E., Walsh, B. T., & Stern, Y. (2006). Set shifting deficit in anorexia nervosa. *Journal of the International Neuropsychological Society, 12*, 431-435.

Tchanturia, K., Harrison, A., Davies, H., Roberts, M., Oldershaw, A., Nakazato, M., et al. (2011). Cognitive flexibility and clinical severity in eating disorders. *Plos One, 6*, e20462.

Tchanturia, K., Morris, R. G., Surguladze, S., & Treasure, J. (2002). An examination of perceptual and cognitive set shifting tasks in acute anorexia nervosa and following recovery. *Eating & Weight Disorders, 7*, 312-315.

Tenconi, E., Santonastaso, P., Degortes, D., Bosello, R., Titton, F., Mapelli, D., et al. (2010). Set-shifting abilities, central coherence, and handedness in anorexia nervosa patients, their unaffected siblings and healthy controls: Exploring putative endophenotypes. *The World Journal of Biological Psychiatry : The Official Journal of the World Federation of Societies of Biological Psychiatry, 11*, 813-823.

Wagner, A., Aizenstein, H., Venkatraman, V. K., Fudge, J., May, J. C., Mazurkewicz, L., Frank, G. K., et al. (2007). Altered reward processing in women recovered from anorexia nervosa. *The American Journal of Psychiatry*, *164*, 1842–1849. doi:10.1176/appi.ajp.2007.07040575

Zastrow, A., Kaiser, S., Stippich, C., Walther, S., Herzog, W., Tchanturia, K., et al. (2009). Neural correlates of impaired cognitive-behavioral flexibility in anorexia nervosa. *The American Journal of Psychiatry, 166*, 608-616.

Appendix: The COVIS Explicit System

COVIS assumes that when learning about new categories, people initially rely almost exclusively on their explicit system. In information-integration tasks, the explicit system fails to find the optimal strategy, so COVIS predicts that under these conditions, control is gradually passed to the procedural system. In rule-based tasks however, the explicit system succeeds in finding the optimal strategy, so COVIS predicts that the procedural system contributes almost nothing in rule-based tasks. Because the tasks used in this article are rule-based, this appendix only describes the COVIS explicit system. For a detailed description of the procedural system, and the system switching algorithm, see Ashby et al. (2011).

In the present application, the COVIS explicit system investigates four possible explicit rules – one-dimensional rules on each of the four stimulus dimensions. Denote the set of these four rules by $\boldsymbol{R} = \{R_1, R_2, R_3, R_4\}$. On each trial, the model selects one of these rules for application by following the algorithm described below. Denote the coordinates of each stimulus on the 4 dimensions by $\underline{x} = (x_1, x_2, x_3, x_4)$. Since the dimensions are binary, each $x_i = +1$ or -1. On trials when the active rule is R_i, a response is selected by using the following decision rule:

$$\text{Respond A on trial } n \text{ if } x_i < \varepsilon\text{; respond B if } x_i > \varepsilon,$$

where ε is a normally distributed random variable with mean 0 and variance σ_E^2. The variance σ_E^2 increases with trial-by-trial variability in the subject's perception of the stimulus and memory of the decision criterion (i.e., perceptual and criterial noise).

Suppose rule R_i is used on trial n. Then the rule selection process proceeds as follows. If the response on trial n is correct, then rule R_i is used again on trial $n + 1$ with probability 1. If the response on trial n is incorrect, then the probability of selecting each rule in the set $\boldsymbol{R}$ for use on trial $n + 1$ is a function of that rule's current weight. The weight associated with each rule is a function of initial bias, the reward history associated with that rule during the current categorization training session, the tendency of the participant to perseverate, and the tendency of the participant to select unusual or creative rules. These factors are all formalized in the following way.

Let $Z_k(n)$ denote the salience of rule R_k on trial n. Therefore, $Z_k(0)$ is the initial salience of rule R_k, which in the present applications were all set equal. The salience of each rule is adjusted after every trial on which it is used, in a manner that depends on whether or not the rule is successful. For example, if rule R_k is used on trial $n - 1$ and a correct response occurred, then

$$Z_k(n) = Z_k(n - 1) + \Delta_C,$$

where Δ_C is some positive constant. If rule R_k is used on trial $n - 1$ and an error occurs, then

$$Z_k(n) = Z_k(n - 1) - \Delta_E,$$

where Δ_E is also a positive constant. The numerical value of Δ_C depends on the perceived gain associated with a correct response and Δ_E depends on the perceived cost of an error.

The salience of each rule is then adjusted to produce a weight, Y, according to the following rules.

1) For the rule R_i that was active on trial *n*,

$$Y_i(n) = Z_i(n) + \gamma,$$

where the constant γ is a measure of the tendency of the participant to perseverate on the active rule, even though feedback indicates that this rule is incorrect (e.g., if γ is small, then switching will be easy, whereas switching is difficult if γ is large).

2) Choose a rule at random from ***R***. Call this rule R_j. The weight for this rule is

$$Y_j(n) = Z_j(n) + \mathbf{X},$$

where **X** is a random variable that has a Poisson distribution with mean λ. Larger values of λ increase the probability that rule R_j will be selected for the next trial, so λ is called the selection parameter.

3) For any other rule R_k (i.e., $R_k \neq R_i$ or R_j),

$$Y_k(n) = Z_k(n).$$

Finally, rule R_k (for all *k*) is selected for use on trial *n* + 1 with probability

$$P_{n+1}(\mathrm{R_k}) = \frac{\mathrm{Y_k}(n)}{\sum_{s=1}^{m} \mathrm{Y}_s(n)}.$$

Table 1. Demographic and clinical information for AN-WR and CW groups.

	AN-WR (n=19)		CW (n=35)		*d*
	M	*SD*	*M*	*SD*	
Age at Study (years)	29.7	6.6	27.7	5.1	0.34
Age at Illness Onset (years)	14.5	2.6	---	---	---
Age Weight Restored (years)	23.3	5.4	---	---	---
Duration Low Weight (years)	8.8	6.5	---	---	---
Duration Weight Restored (years)	6.5	6.5	---	---	---
Current Body Mass Index (kg/m^2)	21.2	1.3	21.9	1.7	0.46
Lowest Body Mass Index (kg/m^2)	14.9	2.8	---	---	---
Change in Body Mass Index (kg/ m^2)	6.3	3.1	---	---	---
TCI					
Novelty Seeking	17.0	6.3	19.9	5.6	0.24
Harm Avoidance	13.7**	7.2	8.4	3.6	1.03
Reward Dependence	16.7	4.4	18.1	3.0	0.37
Persistence	6.4**	2.0	4.8	1.8	0.84
EDI-2					
Drive for Thinness	11.6**	8.0	0.8	1.8	1.86
Bulimia	2.0**	2.1	0.2	0.9	1.11
Body Dissatisfaction	14.3**	8.5	1.8	3.1	1.95
Ineffectiveness	9.2**	9.7	0.3	0.9	1.30
Perfectionism	8.0**	4.8	3.6	3.2	1.08

Notes: TCI = Temperament and Character Inventory subscales; EDI-2 = Eating Disorder Inventory-2 subscales
* p value <0.05, ** p value <0.01

Table 2. Demographic and clinical correlates of Learning Slope and Shift Cost values for the AN-WR and CW groups. Values are Pearson correlations and 95% confidence intervals are in parentheses.

	AN-WR (n=19)		CW (n=35)	
	Learning Slope	Shift Cost	Learning Slope	Shift Cost
Age at Study (years)	-.14 (-.56-.34)	-.16 (-.57-.32)	**.36*** (.03-.62)	-.17 (-.17-.48)
Age at Illness Onset (years)	.09 (-.38-.52)	.12 (-.35-.54)	---	---
Age Weight Restored (years)	.34 (-.14-.69)	-.18 (-.59-.30)	---	---
Duration Low Weight (years)	.25 (-.23-.63)	-.20 (-.60-.28)	---	---
Duration Weight Restored (years)	**-.49*** (.05-.77)	-.03 (-.48-.43)	---	---
Current Body Mass Index (kg/m^2)	-.31 (-.67-.17)	.25 (-.23-.63)	-.04 (-.30-.37)	-.22 (-.52-.12)
Lowest Body Mass Index (kg/m^2)	.43 (-.03-.74)	-.10 (-.53-.37)	---	---
Change in Body Mass Index (kg/ m^2)	**-.52*** (-.79- -.09)	.20 (-.28-.60)	---	---
TCI-2				
Novelty Seeking	.10 (-.37-.53)	.11 (-.36-.54)	-.15 (-.46-.19)	.06 (-.28-.39)
Harm Avoidance	**.61**** (.23-.84)	.17 (-.31-.58)	-.27 (-.55-.07)	-.23
Reward Dependence	-.06 (-.50-.41)	-.15 (-.57-.33)	-.01 (-.34-.32)	.26 (-.08-.55)
Persistence	.16 (-.32-.57)	-.10 (-.53-.37)	.20 (-.14-.50)	.13 (-.21-.44)
EDI-2				
Drive for Thinness	.07 (-.40-.51)	.09 (-.38-.52)	-.07 (-.39-.27)	.07 (-.27-.39)
Bulimia	.16 (-.32-.57)	.00 (-.45-45)	-.07 (-.39-.27)	.23 (-.11-.52)
Body Dissatisfaction	.18 (-.30-.59)	-.01 (-.46-.45)	.03 (-.31-.36)	.14 (-.20-.45)
Ineffectiveness	.13 (-.34-.55)	.09 (-.38-.52)	-.10 (-.42-.24)	.15 (-.19-.46)
Perfectionism	.04 (-.42-.49)	.31 (-.17-.67)	.05 (-.29-.38)	.17 (-.17-.48)

Notes: TCI = Temperament and Character Inventory subscales; EDI-2 = Eating Disorder Inventory-2 subscales

* p value <0.05

Table 3.

Simulation-related parameters in COVIS

Parameters	Adult AN	Adult CW
Δ_e	1.75	0.001
γ	0.8	0.1
λ	1.5	3.9

Table 4.

Other COVIS parameters

Parameters	AN	CW
Δ_c	0.5	-
α	0.8	-
σ_E^2	0.2	-
Θ_{AMPA}	0.001	-
Θ_{NMDA}	0.002	-
D_{slope}	0.8	-
D_{base}	0.2	-
D_{max}	1	-
α_w	0.4	-
β_w	0.19	-
γ_w	0.02	-
σ_p	0.0125	-
Δ_{Oc}	0.004	-
Δ_{Oe}	0.0075	-

Note. See Ashby et al., 2011 and Hélie et al., 2012a, for a full description of COVIS and its parameters.

Figure 1. Example of the category learning task which presents two different sets of computer-generated stimuli of either a cartoon castle or house.

Figure 2. Accuracy (proportion correct) for AN and CW groups. (error bars are standard error of the mean).

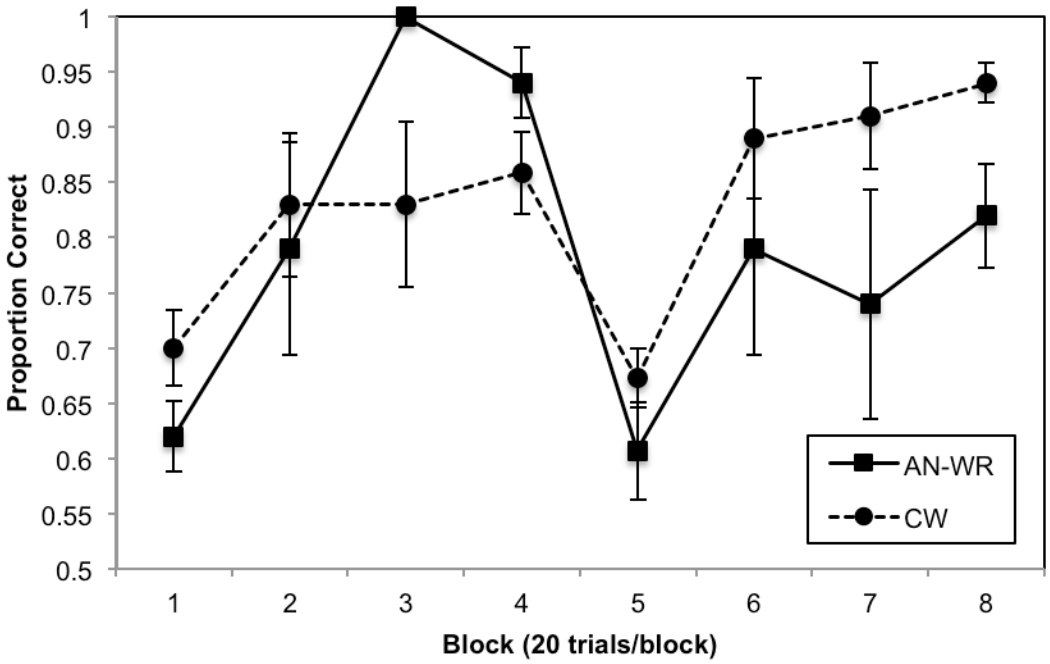

Figure 3. Simulated accuracy (proportion correct) for AN and CW groups.