Memory & Cognition, *in press*

Supervisory Control of Declarative and Procedural Responding

F. Gregory Ashby
University of California, Santa Barbara

Abstract

There is overwhelming evidence that human learning is mediated by multiple systems. If so, then a problem of fundamental importance is to determine how control is passed back-and-forth among these systems. Although more research is needed, existing evidence suggests that system switching does occur, but that it is much less common than assumed by existing models, and that the bias toward strategies mediated by declarative memory is stronger than previously assumed. This shortcoming is addressed by testing a variety of different alternative models of system switching against several empirical benchmarks. The models all assumed separate explicit-rule and procedural/exemplar-similarity categorization systems. The only models that passed all benchmark tests were novel models that incorporate a switch cost and switch systems when explicit-system accuracy drops below a criterion value *and* procedural system accuracy is simultaneously reasonably high. This research suggests a new class of system-switching models, which could be incorporated into any existing multiple-systems model (e.g., COVIS or ATRIUM), and that is much more consistent with existing data than current models.

Keywords: learning; multiple systems; system switching; explicit rules; procedural learning; categorization

1. INTRODUCTION

Learning can be viewed as a process of creating memory traces from past events that affect future performance. Largely for this reason, evidence that humans have multiple memory systems (e.g., Eichenbaum & Cohen, 2001; Squire, 2004; Tulving & Craik, 2005) inspired the development of theories that learning is also mediated by multiple qualitatively distinct systems (Ashby et al., 1998; Erickson & Kruschke, 1998; Hikosaka et al., 1999; Will-

Author Notes. Correspondence concerning this article should be addressed to F. Gregory Ashby, Department of Psychological & Brain Sciences, University of California, Santa Barbara 93106 USA. email: fgashby@ucsb.edu

ingham, 1998). These proposals were generated independently in a diverse set of empirical domains that included categorization (Ashby & Maddox, 2005; Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky & Palmeri, 1998; Nosofsky et al., 1994), identification (Ashby et al., 2001), sequence learning (Hikosaka et al., 1999; Willingham, 1998), language acquisition (Ullman, 2001), and spatial navigation (Packard & McGaugh, 1996). Although these proposals came from different empirical domains, there was widespread agreement that one system uses declarative memory and depends on prefrontal cortex, whereas another uses procedural memory and depends on the striatum.

If humans do have distinct learning systems that use qualitatively different strategies, then logically, there must be some supervisory system or network that determines which of these systems controls behavior at any given point in time. All theories that postulate multiple learning systems must resolve this question, or else no behavioral predictions would be possible. Paul and Ashby (2013) noted that there are at least three different ways that a response could be selected under these conditions, which they called *soft switching*, *hard switching*, and *blending*. Soft switching assumes that trial-by-trial switching between systems is a routine and common occurrence. This is the assumption made by the COVIS and RULEX models of category learning. A hard switch occurs when one system is used exclusively until a single permanent switch is made to the other system. Although this seems like a reasonable assumption, to my knowledge it is not made by any current multiple-systems models. Finally, blending assumes that the outputs of the constituent systems are mixed or blended to produce the final output, and therefore that both systems contribute to performance on every trial. This is the assumption of the ATRIUM model of category learning.

Although countless studies have provided overwhelming evidence for multiple learning systems, only a few have focused on the question of how control is passed back-and-forth between the systems. And if the empirical database on system switching during learning is characterized as meager, then the theoretical development in this area should be considered almost nonexistent. This article addresses this shortcoming by testing a variety of different alternative models of system switching against several empirical benchmarks. A few of these are existing models, whereas most are novel to this article. The end result is a new system-switching model that could be incorporated into virtually any multiple-systems model and that is much more consistent with existing data than any of these original models.

The empirical focus of this article is categorization. As an experimental paradigm, categorization allows high levels of observability into decision processes (Ashby, 2026), and the empirical database, which is enormous, includes most (or all) of the studies that have focused on system switching. In addition, there are prominent and well-established category-learning theories that propose multiple systems and that include computational models of system switching that are straightforward to implement and test. The most prominent multiple-systems models of category learning are COVIS (Ashby et al., 1998, 2011), ATRIUM (Erickson & Kruschke, 1998), and RULEX (Nosofsky & Palmeri, 1998; Nosofsky et al., 1994). All three models assume one system learns explicit rules, and they differ somewhat on the nature of the second system. RULEX assumes that participants memorize exceptions to the rule,[1] ATRIUM assumes the second system categorizes by computing

[1]It is not clear that RULEX actually proposes multiple systems. COVIS and ATRIUM both explicitly assume that the rule system uses a form of declarative memory, whereas the other system uses a form of

exemplar similarities, and COVIS assumes the second system uses procedural learning to acquire stimulus-response (SR) associations.[2]

Scores of empirical studies provide overwhelming support for multiple category-learning systems (for reviews, see e.g., Ashby & Maddox, 2005; Ashby & Valentin, 2017; Ashby et al., 2020; Minda et al., 2024; E. E. Smith & Grossman, 2008). The evidence strongly suggests that declarative memory is used to apply rules and test explicit hypotheses about category membership and procedural memory is used to form many-to-one SR associations of the type assumed by COVIS.[3] Furthermore, the evidence is good that these systems compete against each other for control of responding (e.g., Poldrack et al., 2001).

When COVIS, ATRIUM, and RULEX were first formulated, there were virtually no data that could be used to model system switching. As a result, in each of these models, the assumptions about which system would control responding were made with little or no empirical justification. Although only a few of the thousands of empirical category-learning studies that have been published since these theories were first proposed have focused on this question, fortunately, some results are now known. The most relevant are reviewed in detail in section 3, but briefly, there is now good evidence that system switching does occur, but that it is much less common than assumed by any existing model. Furthermore, there is also evidence that the bias toward strategies mediated by declarative memory is stronger than previously assumed.

2. OVERALL APPROACH

The goal of this article is to test a variety of different models of system switching on qualitative, rather than quantitative empirical benchmarks. This approach avoids the problem of identifying any single individual data sets as gold standards that the various models should be able to fit. Instead, the benchmark tests will examine the ability of the various models to account for qualitative results that characterize many participants in experiments that used certain specific experimental designs. As we will see, this qualitative approach is strong enough to eliminate many alternative models.

Any test of models that describe how participants switch between learning systems must be tested on data in which system switching occurs or is predicted to occur. Obviously, we can learn nothing about system switching by studying tasks in which we know beforehand that all participants will use the same categorization strategy on every trial. There is overwhelming evidence that humans are heavily biased toward simple explicit rules. For example, in free sorting and unsupervised categorization tasks in which no trial-by-trial feedback is provided, virtually all participants use some explicit rule on every trial, regardless of category structure (e.g., Ashby et al., 1999). Similarly, in rule-based (RB) category-learning tasks, in which a simple explicit rule maximizes accuracy, the evidence also suggests that almost all participants persist with explicit rules throughout the task (e.g., Ashby & Maddox,

nondeclarative memory. In contrast, RULEX assumes that applying a rule and memorizing an exception both rely on declarative memory.

[2]The original version of COVIS assumed that the nondeclarative system learned a decision bound of arbitrary slope and intercept. However, Ashby and Waldron (1999) presented empirical evidence that strongly contradicted this assumption. As a result, all subsequent versions of COVIS assumed the nondeclarative system uses procedural learning to acquire SR associations.

[3]There is also more limited evidence that a form of prototype abstraction is mediated by the perceptual representation memory system (Blank & Bayer, 2022; Casale & Ashby, 2008).

2005). Therefore, little can be learned about system switching by studying free sorting, unsupervised, or RB categorization tasks. For these reasons, the benchmark tests are all based on some type of information-integration (II) category-learning task (Ashby & Ell, 2001). In an II task, maximum accuracy requires that information from two or more incommensurable stimulus dimensions is integrated perceptually at a pre-decisional stage. This creates categories that cannot be separated by a simple, explicit rule. Some participants do persist with an explicit rule in II tasks, but dozens of studies have reported evidence that II tasks often recruit procedural learning, whereas RB tasks recruit explicit, rule learning (e.g., Ashby & Valentin, 2017; Ashby et al., 2020). In fact, the evidence suggests that in II tasks successful participants begin with an explicit strategy and then eventually switch to a procedural strategy. For this reason, all benchmark tests will compare performance of the various models to human performance in some II task in which immediate trial-by-trial feedback was provided. The next section describes the empirical benchmark tests that any valid model of system switching must pass. As we will see, the tests include tasks in which virtually all participants successfully switch to a procedural strategy (i.e., benchmark 1), as well as tasks in which most participants fail to switch away from an explicit rule, even though a procedural strategy is optimal (i.e., benchmarks 2 and 3). The correct switching model should predict these results across a broad range of its parameter values.

For each test, the overall strategy will be to simulate responses from many hypothetical participants that switch between systems during learning in the manner prescribed by one of a variety of different system-switching models. Those models are described in Section 4. Next, the simulated responses will be evaluated to assess whether the benchmark test is passed or failed – that is, whether the model successfully predicted the learning system that most participants rely on in the task. This judgment will be qualitative rather than quantitative since there is no gold standard set of results I was trying to fit. For example, in benchmark test 1, the vast majority of participants eventually use a procedural strategy. But of the many published studies that have reported results with the benchmark test 1 categories, virtually all have reported that some participants persist with explicit rules throughout the entire training session. No attempt was made to estimate the proportion of the population who persists with explicit strategies with these categories and under the training conditions assumed in benchmark 1. As a result, there is no gold standard number for the predicted proportion of simulated participants that persist with rules or that switch to a procedural strategy. A model will be judged as passing benchmark 1 if this proportion is any value greater than say 80%. Any model that fails a benchmark test will be rejected. As we will see, this process will allow us to reject all current models and converge on a small set of novel models.

The models describe the conditions that precipitate a system switch, but they do not describe how each system selects a categorization response or how each system learns. As a result, the explicit and procedural systems used in all simulations will exhibit no learning and instead apply the same decision strategy on every trial. So the only simulated learning will be of which system to use. The models will be evaluated by testing whether they correctly predict which categorization system will dominate performance after all learning is complete.

As mentioned above, the evidence strongly suggests that one system learns explicit rules and one system learns SR associations. Therefore, the models will all generate re-

sponses in a manner that is consistent with this evidence. Even so, there is no need to model these categorization responses at the process level. Instead, it suffices to produce responses that would be characterized as having been produced by whichever of the two systems actually generated the response. This can be done by using decision-bound modeling (Ashby, 1992, 2026). Decades of research have established that when people use a simple (one-dimensional) rule, the best-fitting decision bound model uses either a vertical or horizontal decision bound, whereas procedural strategies produce responses that are best separated by a decision bound that is similar to the optimal bound, at least in cases where the optimal bound is linear or quadratic (for a review, see e.g., Ashby, 2026). So for all models, rule-based responses will be generated by appealing to a vertical decision bound and procedural responses will be generated by appealing to a diagonal bound. The models will therefore differ only in the assumptions they make about how and when control is passed back and forth between the rule and procedural-learning systems.

Next, following standard practice, two different decision-bound models will be fit to the last 100 trials of each simulated data set – a one-dimensional (1D) rule model and the general linear classifier (GLC). The 1D rule model assumes a decision bound that is either vertical or horizontal and is compatible with an explicit-rule strategy, whereas the GLC assumes a linear bound of arbitrary slope and intercept and is compatible with a procedural-learning strategy. It is important to stress here that the procedural-learning system does not learn a decision bound of any type (Ashby & Waldron, 1999; Casale et al., 2012). Therefore, the GLC does not provide a process interpretation of procedural learning. Even so, decades of research has established that the GLC provides a better fit than the 1D rule model to data when responding is primarily mediated by procedural learning. So the GLC is used here as a diagnostic tool, rather than as a psychological model of procedural learning. The results of the decision-bound modeling will then be used to assess which learning system each model predicts will dominate the last 100 categorization trials in each benchmark test, and finally these predictions will be tested against the published empirical results.

3. THE EMPIRICAL BENCHMARKS

The best benchmark tests will present a range of problems for the alternative models. I chose three benchmarks that each highlight a different challenge. They all come from II categorization tasks, so in all cases the optimal strategy is procedural. The models all predict that people will begin these tasks by experimenting with simple explicit rules, and as already mentioned, this assumption is strongly supported by the available literature. So in all three benchmarks, optimal performance requires switching from rule-based to procedural responding. In benchmark 1 the overwhelming majority of human participants eventually switch to a procedural strategy, whereas in benchmarks 2 and 3, most participants do not switch and instead persist with a simple rule throughout training. In benchmark 2 the performance of the best rule and the best procedural strategy are both poor, whereas in benchmark 3 both strategies yield good performance. So benchmarks 2 and 3 present the models with different challenges.

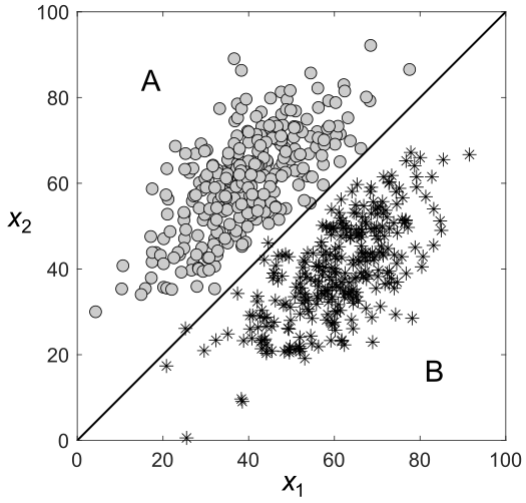

Figure 1 *The II categories used for benchmark test 1. The widely reported empirical result is that most participants in this task will adopt a procedural strategy. Optimal accuracy is 99.7% correct, whereas the accuracy of the best one-dimensional rule is 82% correct.*

3.1 Benchmark 1: II Categories with Little or No Overlap

Scores of studies from many different labs have used an II task in which optimal accuracy is 95% correct or higher and in which the accuracy of the best one-dimensional rule is at least 10% lower than this optimal value. An example is shown in Figure 1, which depicts the coordinates of 300 exemplars from two contrasting II categories denoted A and B that vary on two stimulus dimensions. In this case, the optimal strategy is to respond A to any stimulus falling above the diagonal boundary and B to any stimulus falling below. This strategy correctly categorizes 598 of the 600 stimuli shown in Figure 1. In contrast, the most accurate one-dimensional rule yields an accuracy of 82% correct.

Many studies have reported results with an II task, like the one shown in Figure 1, in which the stimulus dimensions were incommensurable and immediate feedback was provided after every response. Virtually all of these reported that the vast majority of participants eventually adopted a procedural strategy (as just one example, see e.g., Hélie et al., 2010). The evidence for this is far too vast to review here (for reviews, see e.g., Ashby, 2026; Ashby & Valentin, 2017; Ashby et al., 2020). Much of the evidence comes from studies that compared performance in II tasks, like the one shown in Figure 1, with a matched RB task that used identical stimuli and categories, except rotated $\pm 45^\circ$ in stimulus space, so that the optimal category bound was vertical or horizontal. Many results from studies using this design reported evidence that explicit rule learning dominates in the RB task, whereas procedural learning dominates in the II task. Three illustrative examples will be mentioned here. First, humans and monkeys learn the RB task more quickly than the (rotated) II task, whereas pigeons and rats learn the two tasks at exactly the same rate (e.g., Broschard et al., 2019; J. D. Smith et al., 2010, 2011, 2012). These results are consistent with the hypothesis that pigeons and rats lack the well-developed explicit-rule system of humans, and therefore must learn both tasks using an associative learning system similar to human procedural

learning. Second, learning in the II task is impaired with feedback delays as short as 2.5 s, whereas learning in the RB task is unaffected by feedback delays as long as 10 s (Maddox et al., 2003; Maddox & Ing, 2005). These results are relevant because short feedback delays are known to impair procedural learning, but not learning that uses declarative memory (e.g., Foerde & Shohamy, 2011). Third, a simultaneous dual task that requires working memory and executive attention interferes much more with the RB task than with the II task, despite the overall slower initial learning in the II task (Crossley et al., 2016; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006). These results were all predicted a priori by the COVIS theory of category learning, which assumes anatomically and functionally separate explicit-rule and procedural learning systems. Furthermore, there is no single-system account of these many results. For example, there is overwhelming evidence that these qualitative differences are not because the II task is somehow more difficult than the RB task (Ashby et al., 2020).

Assuming that participants are initially biased to begin categorization with some explicit rule, and that most participants will switch to a procedural strategy in the Figure 1 II task, the first benchmark is therefore to test whether the various models successfully predict a switch to some procedural strategy by the end of training in the Figure 1 II task. I will classify a model as passing benchmark 1 if the last 100 trials of at least 80% of its simulated participants are better fit by a model that assumes a procedural strategy than by a model that assumes a simple explicit rule.

3.2 Benchmark 2: II Categories with High Overlap

Benchmark 1 uses a task in which most participants successfully switch to a procedural strategy. Benchmarks 2 and 3 use tasks where most participants fail to switch from an explicit rule. Switch failures of this type are less common in the literature, so benchmarks 2 and 3 each focus on one specific study. Ell and Ashby (2006) examined how category overlap affects the use of explicit versus procedural strategies during II category learning. They reported the results of a number of experiments and conditions, but the most relevant result here is that when category overlap was high, the responses of approximately 70% of participants were best accounted for by a decision-bound model that assumed a one-dimensional explicit rule. The categories they used in these conditions and the optimal decision bound are described in Figure 2. The experiment included four training sessions of 600 trials each, so each participant completed 2400 trials. Optimal accuracy is 69% correct in this task, whereas the accuracy of the best one-dimensional rule is 56% correct. Benchmark 2 will therefore test whether the various models correctly predict that most participants will persist with an explicit-rule strategy under these conditions. A model will be classified as passing benchmark 2 if the last 100 trials of at least 55% of its simulated participants are better fit by a model that assumes a simple explicit rule than by a model that assumes a procedural strategy.

The empirical database that defines benchmark test 2 is considerably smaller than the databases for benchmarks 1 (which is enormous) or 3. Ell and Ashby (2006) reported the results of two experiments that used the Figure 2 II categories (run in two different labs) that differed in the number of training trials. The experiments yielded identical conclusions, but even so, the total number of participants in these two experiments was small (i.e., 10). Although Ell and Ashby (2006) reported that 70% of participants in these two experiments

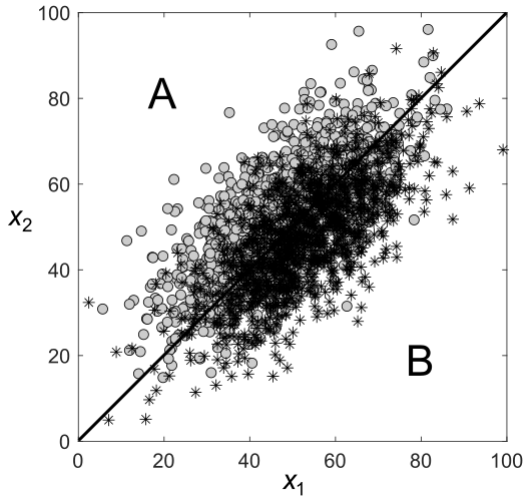

Figure 2 *The II categories used for benchmark test 2. The empirical result is that when II categories are highly overlapping, most participants persist with explicit rule strategies, despite the poor performance they offer. The diagonal line depicts the optimal decision bound. Optimal accuracy is 69% correct, whereas the accuracy of the best one-dimensional rule is 56% correct.*

persisted with a simple rule throughout training, the small sample size suggests that the value 70% might not be exactly representative of a larger population. For this reason, I chose the more lax criterion of 55% for judging whether a model passed or failed benchmark 2.

It is also important to note that although benchmark 2 is based on a small sample size, there is other evidence supporting the empirical phenomenon encapsulated in benchmark test 2 – namely, that when the accuracy of the best procedural strategy and the best rule are both low, people will frequently persist with a simple rule. First, Ell and Ashby (2006) reported results from five different levels of category overlap. The condition shown in Figure 2 was the most extreme. In all conditions in which optimal accuracy was imperfect, the percentage of participants who persisted with a rule increased with category overlap. Second, Ashby and Vucovich (2016) reported the results of several experiments that examined how category overlap affects II category learning. In each experiment, two conditions were matched on optimal accuracy, but in one the discrepant stimuli – that is, the stimuli that would receive an incorrect response from an ideal observer – extended further into the response region of the contrasting category than in the other condition. In all experiments, the use of procedural strategies was lower in the conditions where the categories overlapped over a greater area. Third, note that benchmark test 2 is consistent with results of unsupervised II tasks. When there is no feedback that produces high reward rates for procedural responding, the evidence suggests that all II participants resort to a simple one-dimensional rule (Ashby et al., 1999).

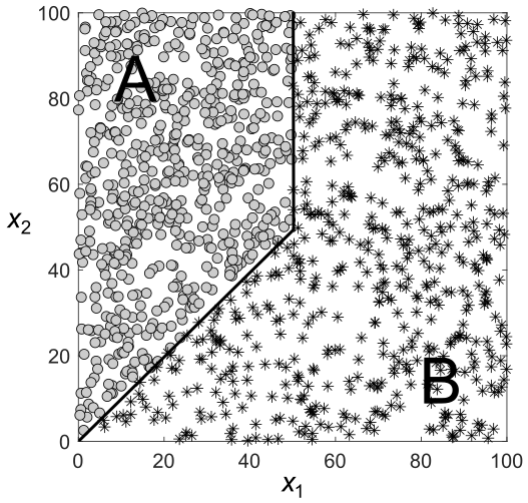

Figure 3 *The hybrid categories used for benchmark test 3 along with the optimal decision bound. The empirical result is that almost all participants persist with a simple one-dimensional rule in this task. Optimal accuracy is 100% correct, whereas the accuracy of the best one-dimensional rule is 90% correct. The accuracy of the diagonal linear bound* $x_2 = x_1$ *is also 90% correct.*

3.3 Benchmark 3: Hybrid Categories

In one of the few studies with the principal goal of studying trial-by-trial system switching, Ashby and Crossley (2010) constructed the hybrid categories shown in Figure 3. Each category contained 600 stimuli and all participants completed the 1200 categorization trials in two separate 600-trial sessions. Note that for these categories a one-dimensional explicit rule is optimal when the stimulus value x_2 is large and a procedural strategy is optimal when x_2 is small. Therefore, if trial-by-trial system switching is routine, as assumed by COVIS, then most participants should respond with near optimal accuracy in this task (i.e., 100% correct). In contrast to this prediction, only 2 of 53 participants appeared to adopt a trial-by-trial switching strategy of this type. The most prevalent strategy, by far, was a simple one-dimensional rule. It is also important to note that both COVIS and ATRIUM predict that the procedural system should easily be able to learn the nonlinear optimal bound shown in Figure 3.[4] As a result, these theories predict that even a single hard switch to procedural responding should result in near optimal accuracy. Benchmark 3 is therefore to test whether the various models will predict that participants persist with some explicit-rule strategy while learning the hybrid categories of Figure 3. A model will be classified as passing benchmark 2 if the last 100 trials of at least 60% of its simulated participants are better fit by a model that assumes a simple explicit rule than by a model that assumes a procedural strategy.

It is also worth noting that other results support the key premise of benchmark test 3 – namely, that people will persist with rules in an II task in which the best procedural

[4]To be clear, neither model assumes participants actually learn a decision bound. Rather they both assume that the decision bound is just the set of points that separates stimuli that the models learn to associate with contrasting category responses.

Table 1 *Models tested in this article. E = explicit system; P = procedural system; θ = COVIS system weight; N = last N trials.*

		Switching		
		hard	**soft**	**blending**
Switch Costs	**none**	hard-COVIS	COVIS	ATRIUM
	switches only when E accuracy is low	hard-E-θ	soft-E-θ	
		hard-E-N	soft-E-N	
	switches only when E accuracy is low & P accuracy is high	hard-EP-θ	soft-EP-θ	
		hard-EP-N	soft-EP-N	

strategy only slightly outperforms the best rule. For example, Ashby et al. (1998) reported the results of an II experiment in which the best procedural strategy only outperformed the best rule by about 3%. There were two conditions. In one, the slope of the optimal decision bound was steeper than the diagonal bound that is optimal in benchmark test 1, and in the other condition the optimal bound was shallower than this. Despite practicing these tasks for 2000 trials distributed over four experimental sessions, the responses of many participants during the final session were best described by a simple one-dimensional rule.

4. ALTERNATIVE MODELS OF SYSTEM SWITCHING

The goal of this article is to investigate alternative models of system switching, rather than alternative models of how the various putative systems learn. In all three benchmark tests, the strategy analysis that decided whether each participant was using an explicit-rule or procedural strategy was based on the responses collected during the final trials of the task – typically the final 100 trials. In most cases, this is after all learning has been completed. Therefore, in each benchmark test, I assumed that each system used its best strategy on every trial and that the only learning that occurred was of which system should control responding on each trial.

The various models that were tested are described in Table 1. A detailed description of each model in provided below, but briefly, they differed on three factors. First, the models differed on whether they assumed a hard switch, a soft switch, or blending. Second, they differed on whether they assumed a switch cost, and if they did assume a switch cost, how this was computed. Finally, they differed in how they monitored performance in each system. Some models used the COVIS system weights θ_{E} and θ_{P} to monitor performance, and some used the number of categorization errors that the system made in the last N_t trials.

4.1 COVIS

Virtually all existing models of system switching assume no switch costs and that switching on each trial is optimal or nearly optimal – that is, control is given to the system that is judged to be best prepared for that trial (e.g., most likely to respond correctly). All such models assume soft switching because theoretically, control could pass back and forth between the two systems on virtually every trial. At the highest level, the various models are fundamentally different, primarily because they make different assumptions about the nature of each system. Even so, at the system-switching level, they differ only slightly (e.g., they might differ in what statistic they assume is optimized). COVIS is in this class of models, as is RULEX (Nosofsky et al., 1994) and the multiple-systems reinforcement learning model of Daw et al. (2005). When embedded into the rule and procedural categorization models used in all simulations reported in this article, however, all such models should make nearly identical predictions – that is, they should all have identical outcomes on the three benchmark tests.

COVIS will be used as an example model of this class. COVIS assumes that on each trial, control of responding is given to the system with the greatest weighted confidence in the accuracy of the response it recommends, where the weight is determined by the prior success rate of that system. In both COVIS systems, response confidence is based on the distance from the current percept to the decision bound that describes the system's response strategy. For example, if the explicit system is using a one-dimensional rule on dimension 1, then its confidence on a trial when the percept has coordinates (x_1, x_2) equals

$$H_{\mathrm{E}} = |x_1 - X_{\mathrm{C}_1}|, \tag{1}$$

where X_{C_1} is the value of the response criterion on dimension 1.

The current version of the COVIS procedural system does not learn a decision bound. Rather it learns to assign a response to each region of perceptual space. Even so, one could operationally define a decision bound as the set of points that separate percepts associated with contrasting responses. Given this definition, response confidence equals the absolute value of the distance from the percept to this operationally defined decision bound. Call this distance H_{P}.

The success rate of each system is updated trial-by-trial depending on whether the response recommended by the system was correct or incorrect. Specifically, after feedback is provided on trial n, the system weights θ_{E} and θ_{P} for the explicit and procedural systems, respectively, are adjusted up if the suggested response of the system was correct according to the difference equation

$$\theta_{\mathrm{J}}(n+1) = \theta_{\mathrm{J}}(n) + \Delta_{+}[1 - \theta_{\mathrm{J}}(n)], \tag{2}$$

where J = E or P and Δ_{+} is a positive constant. The $[1 - \theta_{\mathrm{J}}(n)]$ multiplier prevents θ_{J} from exceeding 1.0. Similarly, if the system recommended an incorrect response then its associated weight is decreased by

$$\theta_{\mathrm{J}}(n+1) = \theta_{\mathrm{J}}(n) - \Delta_{\text{-}}\theta_{\mathrm{J}}(n), \tag{3}$$

where J again equals E or P and $\Delta_{\text{-}}$ is a positive constant. Now the multiplier $\theta_{\mathrm{J}}(n)$ prevents θ_{J} from dropping below 0. Δ_{+} and $\Delta_{\text{-}}$ typically take values in the range [.01, .20].

To reflect an initial bias towards rules, the initial values of θ_{E} and θ_{P} are set to $\theta_{\mathrm{E}}(0) = .99$ and $\theta_{\mathrm{P}}(0) = .01$.

Given these definitions, on each trial COVIS uses the decision rule:

Emit the response suggested by the explicit system if $\theta_{\mathrm{E}} H_{\mathrm{E}} \geq \theta_{\mathrm{P}} H_{\mathrm{P}}$;
Otherwise emit the response suggested by the procedural system.

Note that this rule theoretically allows for system switching on every trial.

It turns out that Equations 2 and 3 cause θ_{J} to oscillate around the true probability that system J responds correctly on any given trial. To see why this is true, suppose $\Delta_+ = \Delta_- = 1/n$ in Equations 2 and 3. Now let $X_i = 1$ if the response on trial i was correct and $X_i = 0$ if the response was incorrect. Then note that Equations 2 and 3 reduce to the single equation

$$\theta_{\mathrm{J}}(n+1) = \theta_{\mathrm{J}}(n) + \frac{1}{n}[X_n - \theta_{\mathrm{J}}(n)]. \tag{4}$$

Now suppose $\theta_{\mathrm{J}}(0) = 0$. Then it is straightforward to show that Equation 4 reduces to the sample mean[5]

$$\theta_{\mathrm{J}}(n+1) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{5}$$

In other words, the system weight on trial $n+1$ is set to the mean accuracy on the preceding n trials. So under these conditions, $\theta_{\mathrm{J}}(n)$ is the optimal estimator of system J categorization accuracy. For this reason, COVIS tends to weight each system by its categorization accuracy.

It is also straightforward to investigate a version of COVIS that assumes a hard switch, which we can identify as hard-COVIS. This version is identical to COVIS except that after the first switch to the procedural system, no more switches occur, and instead all future responses are made with the procedural system. Note that this model predicts more procedural responding than COVIS. In particular, it predicts that procedural responding will dominate in all tasks in which COVIS makes this prediction, but in addition it predicts that there might also be tasks in which procedural responding dominates that COVIS predicts should elicit mostly rule-based responding.

4.2 ATRIUM

ATRIUM assumes blending of the outputs of two separate category-learning systems – one that uses explicit rules and one that responds based on exemplar similarities (Erickson

[5]Note that Equation 4 implies that

$$\begin{aligned}\theta_{\mathrm{J}}(n+1) &= \frac{1}{n}X_n + \frac{n-1}{n}\theta_{\mathrm{J}}(n) \\ &= \frac{1}{n}X_n + \frac{1}{n}X_{n-1} + \frac{n-2}{n}\theta_{\mathrm{J}}(n-1).\end{aligned}$$

Iterating out in this fashion produces the result. If $\theta_{\mathrm{J}}(0)$ is nonzero, then Equation 4 reduces to

$$\theta_{\mathrm{J}}(n+1) = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\theta_{\mathrm{J}}(0).$$

As a result, note that the contribution of the initial value of θ_{J} to its final value decreases as n increases.

& Kruschke, 1998). In a task with two categories A and B, ATRIUM assumes that the probability of responding A is a weighted sum of the probability that the exemplar-similarity system alone would respond A plus the probability that the rule system alone would respond A. Specifically, let $P_{\mathrm{E}}(A|\underline{\mathrm{x}})$ denote the probability that the exemplar-similarity system responds A to a percept with coordinates $\underline{\mathrm{x}} = (x_1, x_2)$ and $P_{\mathrm{R}}(A|\underline{\mathrm{x}})$ denote the analogous probability for the explicit-rule system. Then ATRIUM assumes that the probability the participant responds A on this trial equals

$$P(A|\underline{\mathrm{x}}) = \alpha P_{\mathrm{E}}(A|\underline{\mathrm{x}}) + (1-\alpha)P_{\mathrm{R}}(A|\underline{\mathrm{x}}), \tag{6}$$

where α is a parameter constrained to the interval [0, 1]. Note that any two-system model that assumes blending must satisfy this same equation, where the subscripts E and R denote the two systems assumed by the model. As a result, although I will continue to refer to this model as ATRIUM, it is important to note that in the present applications, it makes identical predictions to any generic model that assumes blending.

The parameter α is adjusted through learning to reflect the categorization success of the exemplar-similarity system, relative to the rule system. Although this learning process is complex in the sense that it is described by a number of free parameters, in the present applications, α would be constrained to some intermediate value because the ratio of the accuracy of the best exemplar-similarity strategy to the best rule strategy is never greater than 1.25 in any benchmark test. As a result, ATRIUM predicts that both systems will make a significant contribution to responding in all three benchmark tests.

At the mathematical level, exemplar models of categorization are a special case of the COVIS procedural-learning model (Ashby & Rosedahl, 2017). Exemplar models assume categorization decisions are based on the summed similarity of the presented stimulus to memory representations of all previously seen stimuli (e.g., Nosofsky, 1986). The COVIS procedural-learning model assumes that each categorization trial alters the strength of synapses between visual cortex and the striatum. It turns out that, under certain simplifying assumptions, the summed similarities computed by the exemplar model are exactly equal to the strengths of these cortical-striatal synapses. As a result, in all three benchmark tests, the two systems assumed by ATRIUM make essentially equivalent predictions to the two systems assumed by COVIS, even though the COVIS procedural system and the ATRIUM exemplar-similarity system make very different psychological assumptions. Therefore, in the present applications, the only real difference between COVIS and ATRIUM is that COVIS assumes soft switching, whereas ATRIUM assumes blending.

Technically, ATRIUM is not a multiple-systems model, nor is any other model that assumes blending – at least not according to the definition proposed by Ashby and Bamber (2022), who formalized a definition of a system that is popular in the memory literature. Multiple memory systems are often identified by lesion experiments. If there is a single system, then a lesion anywhere will impair performance in all tasks. In contrast, if there are multiple systems then a lesion to one system will impair performance on all tasks that recruit that system, but will have no effect on performance on tasks mediated completely by some other system. For example, lesioning the hippocampus impairs episodic memory but not procedural memory, whereas lesions to the striatum impair procedural memory but not episodic memory (e.g., Packard & McGaugh, 1992). In ATRIUM, because of the blending assumption, both systems contribute to the probability of responding A in all

tasks, so lesioning either system will change performance in all tasks, which is consistent with a single underlying system. Similarly, there is no system switching in ATRIUM because behavior is never controlled by either the exemplar or rule modules. Instead, both modules contribute to responding on all trials. On the other hand, ATRIUM does make an optimality assumption that is similar to the assumptions made by the models considered in the previous section. Specifically, ATRIUM assumes that the contribution of each system is proportional to its historical categorization accuracy.

4.3 Models that Switch Systems When Explicit-System Accuracy is Low

Neither COVIS nor ATRIUM allow for any type of switch cost. For example, COVIS assumes that switching between systems incurs no cost in either effort or performance. In other words, COVIS predicts that the categorization speed and accuracy of each system is the same regardless of which system controlled responding on the previous trial. A large literature has documented that switching from one declarative memory task to another incurs a significant switch cost (e.g., Monsell, 2003), as does switching from one procedural memory task to another (Kita et al., 2025). Furthermore, there is also evidence that the cost of switching from a declarative- to a procedural-memory task, or vice versa, is even larger (Crossley et al., 2018). In all these cases, performance is considerably worse on the first trial after a task switch. Therefore, one obvious revision of existing models of system switching is to allow for some type of system-switch cost.

The analyses of the data that underlie the benchmark tests focused only on asymptotic performance. Since the same task was used on all trials, trial-by-trial switch costs are unobservable in these studies. For this reason, no attempt will be made to model switch costs directly. Even so, it seems plausible that switch costs – no matter how they are computed – will reduce the frequency of system switching. If so, then asymptotic performance could also be affected. Therefore, I investigated several different models that assumed switch costs would reduce the frequency of system switching. These models either assumed a hard switch – that is, a single permanent switch from an explicit rule to procedural responding – or else soft switching, but at a frequency that is considerably lower than predicted by COVIS.

One class of models of this type persists with explicit-rule responding until the accuracy of the explicit system drops below a criterion level. At this point a switch is made to the procedural system. I investigated four different versions of this model by crossing two different levels of two factors (i.e., see Table 1). The first factor was whether or not the switch was permanent – that is, whether the switching was hard or soft. In the versions that assumed a permanent switch, the model either made no switches and persisted with explicit-rule strategies throughout all categorization trials, or else made a single hard switch to the procedural system and then persisted with a procedural strategy for the remaining trials. In the versions that allowed multiple switches, the model used the explicit system on all trials when explicit system accuracy was above the criterion value, and used the procedural system on all trials when explicit system accuracy was below criterion.

The second factor was based on how performance is monitored. One way is to set a criterion on the COVIS explicit system weight θ_{E}. Specifically, in these versions, a switch to the procedural system occurs when $\theta_{\mathrm{E}}(n) < C_{\mathrm{E}}$, where C_{E} is a parameter. The hard- and soft-switch versions of these models are designated as models hard-E-θ and soft-

E-θ, respectively, where the E indicates that the only performance monitoring is of the explicit system. Both models have three free parameters: Δ_+ and Δ_-, as in COVIS, and the criterion for switching C_{E}. As in COVIS, the hard- and soft-switch versions of this model are initially biased toward using the explicit system (i.e., initially $\theta_{\mathrm{E}} = .99$) and a switch from the explicit system to the procedural system occurs when θ_{E} drops below C_{E}. In the hard-switch version of the model, a permanent switch is made on the first trial for which $\theta_{\mathrm{E}}(n) < C_{\mathrm{E}}$, whereas in the soft-switch version, the procedural system is used on all trials for which $\theta_{\mathrm{E}}(n) < C_{\mathrm{E}}$ and the explicit-rule system is used on all trials for which $\theta_{\mathrm{E}}(n) \geq C_{\mathrm{E}}$. The second level of this factor uses response accuracy directly to monitor explicit-system performance, instead of the COVIS system weights. Specifically, in these versions, a switch to the procedural system occurs if the proportion of error trials in the last N_t trials is greater than or equal to β_{E}, where N_t and β_{E} are free parameters. The hard- and soft-switch versions of these models are referred to as models hard-E-N and soft-E-N, respectively.

We saw earlier that the COVIS system weight θ_{E} tends to oscillate around the true probability that the explicit system responds correctly on any given trial. In contrast, the models that use the number of errors in the preceding N_t trials use a highly suboptimal estimate of categorization accuracy. For example, these estimators are not even sufficient statistics for the true accuracy rate. Because they depend only on the preceding N_t trials, they are extremely volatile. As a result, simulated participants from these proportion-of-errors models are more likely to produce responses that would be uncharacteristic of the model if a more powerful statistical estimator was used to judge system performance. Even so, note that the optimality of models hard-E-N and soft-E-N will improve as the value of N_t increases.

Obviously, the hard-switching versions of these models switch between systems much less frequently than COVIS, but this is also true of the soft-switching versions. For example, the COVIS system weights are measures of the historical success of each system and so only change sluggishly from trial to trial. This inertia reduces the likelihood of back-and-forth, trial-by-trial system switching. COVIS also uses these weights, of course, but COVIS multiplies each weight on every trial by a measure of how confident the system is in its ability to categorize the current stimulus accurately. These confidence measures fluctuate wildly trial by trial, making this weighted measure much more volatile that the system weights themselves. As a result, all versions of this class of models could be viewed as having built-in switch costs.

Finally, note that it is not possible to construct a version of ATRIUM that includes switch costs, since ATRIUM never switches systems.

4.4 Models that Also Monitor Procedural-System Accuracy

In the models described in the previous subsection, the decision to switch to the procedural system depends only on the accuracy of the explicit system. The performance of the procedural system is never monitored and plays no role in the switch decision. In other words, a switch to the procedural system is made whenever the explicit system performs sufficiently poorly – even if the procedural system performance would be even worse. In contrast to this assumption, there are several reasons to believe that a switch to the procedural system occurs only when procedural-system accuracy is at least as good as explicit-system

accuracy. A complete review is beyond the scope of this article, but briefly, the evidence includes the following. First, category-learning curves do not exhibit a pronounced dip at some intermediate point of learning that should occur if a switch is made to a less accurate categorization system (e.g., J. D. Smith & Ell, 2015). Second, there is no evidence that a switch to the procedural system ever occurs under conditions that impair procedural learning. For example, when no feedback is provided, or the feedback is delayed by a few seconds, all participants in II tasks appear to persist with explicit strategies (e.g., Ashby et al., 1999; Maddox et al., 2003).

For these reasons, it is important to consider models that monitor procedural-system accuracy and switch from explicit-system responding only when there is some guarantee that such a switch will not cause accuracy to drop precipitously. Note that any such model requires that the procedural system learns in the background when responding is controlled by the explicit system. Without such background learning, procedural-system accuracy would remain at chance during explicit-system control, and as a result, a switch would only occur when explicit-system accuracy also drops to chance. In fact, the evidence is good that the procedural system does learn in the background when responding is controlled by the explicit system (Crossley & Ashby, 2015). See the Discussion for more details on this important issue.

The same two factors that were described in the previous subsection were crossed here to create four versions of this general model. Specifically, switching was either hard or soft, and performance was evaluated either via the COVIS system weights or by the number of errors made by each system during the preceding N_t trials. The resulting four models are denoted as models hard-EP-θ and soft-EP-θ for the hard- and soft-switch versions that use the COVIS system weights to monitor performance, respectively, and as models hard-EP-N and soft-EP-N for the hard- and soft-switch versions that decide whether or not to switch based on performance during the last N_t trials, respectively.

Models hard-EP-θ and soft-EP-θ switch to the procedural system either on the first trial (hard-EP-θ) or on every trial (soft-EP-θ) for which $\theta_{\text{E}} < C_{\text{E}}$ and at the same time $\theta_{\text{P}} > C_{\text{P}}$, where C_{E} and C_{P} are free parameters. As in all other models, there is an initial bias toward using the explicit system (i.e., initially $\theta_{\text{E}} = .99$ and $\theta_{\text{P}} = .01$). The full versions of both models have four free parameters: $\Delta_+, \Delta_-, C_{\text{E}}$, and C_{P}. This is more than all other models, and it is important to ensure, for example, that if either model passes the benchmark tests, it is not simply because their extra parameters make them more mathematically flexible. As a result, I will only investigate versions of these two models in which $\Delta_+ = \Delta_- = \Delta$. So these models have three free parameter: Δ, C_{E}, and C_{P} – the same number as the models considered in the previous section. The versions that count the number of recent errors switch to the procedural system either on the first trial (hard switch) or on every trial (soft switch) for which the proportion of explicit-system errors in the last N_t trials is greater than β_{E} and the proportion of procedural-system errors is less than β_{P}, where N_t, β_{E} and β_{P} are free parameters.

The soft-switching versions of this model switch systems less frequently than the soft-switching versions of the models that only monitor explicit system accuracy. This is because a system switch requires two conditions, rather than one – specifically, poor explicit-system and good procedural-system performance. Therefore, this class of models could be viewed as displaying larger switch costs than the models that base the decision to switch

only on explicit-system accuracy.

5. METHODS

For each benchmark test and each model, I simulated the performance of 100 separate hypothetical participants for a variety of different parameter settings. In each case, I randomly shuffled the stimulus presentation order for each hypothetical participant and I added unique normally distributed perceptual noise to each stimulus. In all cases, the noise had mean zero on both dimensions, standard deviation 10, and the noise values on the two dimensions were uncorrelated. Note that in all three benchmark tests, the stimuli all have values in the range [0,100] on both stimulus dimensions. Given this stimulus space, a perceptual noise standard deviation of 10 is representative of perceptual-noise parameter estimates when decision bound models are fit to data from RB and II experiments.

COVIS and ATRIUM both assume that the rule system will learn to apply the most accurate one-dimensional explicit rule in all three benchmark tests, whereas the other system (i.e., procedural or exemplar-similarity) will learn a strategy that mimics the optimal bound. Therefore, in all simulations, I assumed that the model's explicit system always used an explicit rule on dimension 1 that could be described by a vertical decision bound with x_1 intercept equal to 50, and I assumed that the procedural/exemplar-similarity system used a strategy that could be described by a diagonal decision bound with slope one and intercept zero. Note that this diagonal bound is optimal in benchmarks 1 and 2, but not in benchmark 3. With the hybrid categories of benchmark 3, the optimal bound, shown in Figure 3, is nonlinear. Recall that the responses of the vast majority of participants in the benchmark 3 categorization task are compatible with an explicit rule that is best described by a vertical decision bound. The models all learn to increase the control of the more accurate of the two categorization systems. Therefore, the models are more likely to pass benchmark test 3 – that is, to produce responses that are best described by an explicit rule – if the procedural system uses a suboptimal bound. As a result, models that fail under these favorable conditions can be rejected with confidence. For this reason, in all simulations of the benchmark 3 hybrid task, I assumed the procedural system used a linear bound with slope 1 and intercept 0.

Next, I fit two types of decision-bound models to the last 100 trials of each simulated data set – the 1D classifier that is compatible with an explicit-rule strategy and the GLC that is compatible with a procedural-learning strategy. By this point of the categorization session, all learning is presumed to be complete, and so the data are assumed to be stationary. The decision-bound modeling analysis will then identify the type of decision strategy that dominates after all learning is complete. See Ashby (2026) for computational details of each model, as well as for the Matlab code I used to fit each model to the simulated data. Briefly though, the explicit-rule model assumed that participants set a decision criterion on stimulus dimension 1 and responded with the explicit rule: "Respond A if the stimulus has a small value on dimension 1, otherwise respond B". This 1D rule model has two free parameters – a response criterion on dimension 1 and the variance of perceptual noise. There was no reason to fit the analogous model that assumed an explicit rule on stimulus dimension 2. First, the simulated data were always generated under the assumption that the explicit system was using a vertical decision bound. Second, the category structures in benchmark tests 1 and 2 are symmetric, in the sense that the most accurate one-dimensional

rules are equally accurate on the two dimensions. In benchmark 3, a one-dimensional rule on dimension 1 is considerably more accurate than any one-dimensional rule on dimension 2. The GLC, which is compatible with procedural responding, assumes a linear bound that can have any slope and intercept. The GLC has three free parameters – the slope and intercept of the best-fitting bound and the variance of perceptual noise. Note that the 1D model is a special case of the GLC in which the linear bound is constrained to be vertical. The BIC goodness-of-fit measure was used to select the model that gave the best account of each data set.

In summary, for each model and each benchmark test, I simulated responses from 100 hypothetical participants for a variety of different parameter settings. Next, for each set of simulated responses, I fit both the 1D classifier and the GLC to the data from the last 100 simulated trials and compared their fits using BIC. If the 1D model fit better, I classified the simulated participant as using an explicit rule, whereas if the GLC fit better, the participant was classified as using a procedural strategy.

6. RESULTS

6.1 COVIS

For each benchmark test, the performance of 100 participants was simulated for each of nine different combinations of the parameters Δ_+ and Δ_-. Specifically, three values of Δ_+ (i.e., 0.001, 0.101, and 0.201) were crossed with three values of Δ_- (i.e., 0.001, 0.101, and 0.201). Therefore, for each benchmark test, the performance of 900 separate participants was simulated. It is also important to note that the chosen parameter values effectively cover the entire parameter space. First, when either Δ_+ or Δ_- are less than the minimum value of 0.001, there is virtually no learning over the course of the simulated trials. In the absence of any learning, COVIS predicts that the initial bias toward explicit rules will persist throughout the task, and as a result, the model fails benchmark 1. Second, when the values of Δ_+ or Δ_- are set to values substantially larger than the maximum of 0.201, then as is well known in neural network theory, learning becomes unstable.

In benchmark 1, for 8 of the 9 parameter combinations, the GLC fit better than the 1D model for 95 or more of the 100 simulated participants. The one exception occurred when $\Delta_+ = \Delta_- = .001$, in other words, when there was almost no learning about which system should control responding. As a result, the initial explicit rule bias persisted far longer than with any other parameter combination. Even so, the GLC still fit better for 56 of the 100 simulated participants. In other words, for almost all parameter settings, COVIS predicts that procedural strategies will dominate after learning is complete with the benchmark 1 categories, and as a result, COVIS successfully passed benchmark test 1 for almost all parameter values.

COVIS badly failed benchmark test 2. For all nine parameter combinations, the GLC provided a better fit than the 1D rule model for a minimum of 64 simulated participants (when $\Delta_+ = .101$ and $\Delta_- = .001$) and a maximum of 81 (when $\Delta_+ = .001$ and $\Delta_- = .101$). The frequent system switching predicted by the model means that under virtually all learning conditions, there are enough trials when the emitted response is under procedural control that the resulting data are better fit by the GLC than by the 1D model.

Similarly, COVIS also badly failed benchmark test 3. For all 9 parameter combinations, the GLC fit better for 84 or more of the 100 simulated participants, and for 8 of the 9 combinations, the GLC fit better for 94 or more participants. The best performance by the 1D rule model was again when $\Delta_+ = \Delta_- = .001$, which slows system learning to a crawl and makes the initial bias towards rules difficult to overcome. But even in this case, the GLC fit better than the 1D model for 84 of the 100 participants. Furthermore, the difference in fits between the two models was substantial. For example, the mean BIC score across the 900 simulated participants (i.e., 9 parameter combinations $\times$ 100 participants each) was 43.8 for the GLC and 55.8 for the 1D model. So assuming equal prior probabilities that each model is correct, the probability that the GLC is correct (assuming that one of the GLC and 1D models are correct) is about .998 (e.g., Raftery, 1995). Therefore, COVIS fails to predict that most participants will resort to some simple one-dimensional rule when confronted with the hybrid categories of benchmark test 3.

Recall that the COVIS responses in benchmark test 3 were simulated under the assumption that the procedural system used a strategy that could be described by a linear decision bound, rather than by the nonlinear bound that is optimal with these categories and that can be learned easily by the COVIS procedural system (i.e., the striatal pattern classifier; Ashby & Waldron, 1999). Before moving on to the next model, it is worth considering how this design choice might have affected the conclusion that COVIS predicts that procedural strategies will dominate with the benchmark test 3 hybrid categories. The optimal strategy with these categories, which is described by the bound shown in Figure 3, achieves 100% accuracy (i.e., in the absence of perceptual noise), whereas the linear bound that was used to simulate procedural responses and the one-dimensional rule used to simulate explicit-rule responses both achieve accuracies of 90% correct. COVIS is initially biased toward explicit rules. If the procedural system is more accurate than the explicit system, then it can eventually overcome this bias (i.e., by adjusting the system weights via Equations 2 and 3). In the current simulations, the two systems were equally accurate, which was reflected in the final system weights. For 8 of the 9 parameter combinations, the two final system weights differed by less than .015. The only exception occurred when $\Delta_+ = \Delta_- = .001$ and learning was impoverished, in which case the explicit system weight ($\Delta_- = .875$) was considerably larger than the procedural system weight ($\Delta_+ = .580$). If the procedural system responses were generated via the optimal decision bound, then Equation 2 predicts that because of the higher accuracy of the procedural system, the procedural system weights would be greater than in the current simulations. As a result, the prevalence of procedural strategies would be even greater than observed here. So assuming the procedural system used a suboptimal decision strategy biased the test toward explicit rules. Despite this bias, the results overwhelmingly suggest that COVIS incorrectly predicts that procedural strategies should dominate with the hybrid categories of benchmark test 3.

Similarly, the choice to not fit a procedural decision-bound model that used a nonlinear bound also had no effect on the conclusion that COVIS predicts that procedural strategies should dominate in benchmark test 3. If a nonlinear decision-bound model fit worse than the 1D model, then the conclusion would be unchanged since the GLC almost always fit better than the 1D model. If a nonlinear model fit better than the 1D model then our conclusion would be reinforced because there would be even more evidence that procedural strategies dominate with the benchmark test 3 hybrid categories.

It should also be noted that a version of COVIS that assumes a hard switch, rather than a soft switch (i.e., model hard-COVIS; see Table 1), is even easier to reject. The original soft-switch version of COVIS predicts that procedural strategies will dominate in benchmark tests 2 and 3 for all values of its two parameters Δ_+ and Δ_+. It therefore predicts an early switch to the procedural system in both tasks. The hard-switch version predicts that after the first of these switches, procedural strategies will thereafter be used on all trials. So the hard-switch version predicts that procedural responding in benchmarks 2 and 3 must be at least as frequent as in the soft-switch version. As such, the hard-switch version also fails benchmark tests 2 and 3.

6.2 ATRIUM

For each benchmark test, the performance of 100 participants was simulated for 11 different values of the parameter α that spanned the entire parameter space (from $\alpha = 0$ to $\alpha = 1$, in increments of .1). Recall that α, which is restricted to the range [0,1], is the mixing weight on the procedural/exemplar-similarity system. So if $\alpha = 0$ the procedural system is ignored and the categorization response is determined completely by the explicit system, whereas if $\alpha = 1$ the procedural system determines the response and the explicit system is ignored. The results are shown in Figure 4. Note that the results are similar for all three benchmarks. In each case, the proportion of simulated participants for which the responses during the last 100 trials were best accounted for by the GLC increases monotonically with α from zero to one. Recall that the benchmark tests require this proportion to be high in benchmark 1 and low in benchmarks 2 and 3. Note that there is no value of α that passes these tests. For example, in the case of the low-overlap II categories of benchmark 1, a value of $\alpha > .4$ is required to predict that the responses of at least 80% of participants are best fit by the GLC. However, with this value of α, ATRIUM predicts that the GLC should almost always fit better in benchmarks 2 and 3. As a result, ATRIUM badly fails the benchmark tests.

ATRIUM could pass all three benchmarks if the value of α was allowed to vary without constraint across the three tests. For example, the three benchmarks are passed if $\alpha > .4$ in benchmark 1 and $\alpha \leq .1$ in benchmarks 2 and 3. The problem with this scenario is that there is no mechanism in ATRIUM that would account for such different values of α. Equation 6 shows that α is the weight that ATRIUM assigns to the procedural/exemplar-similarity system and $1 - \alpha$ is the weight it assigns to the rule system. As noted earlier, ATRIUM adjusts α based on the categorization success of the procedural/exemplar-similarity system, relative to the rule system. So in tasks where the procedural/exemplar-similarity system is more accurate, α should be greater than .5, whereas in tasks where the rule system is more accurate, α should be less than .5. In benchmark tests 1 and 2, the procedural/exemplar-similarity system outperforms the best rule by at least 13%. As a result, ATRIUM predicts that α should be greater than .5 in these tasks. In benchmark 3, the two systems are equally accurate, so α should be approximately equal to .5. Figure 4 shows that ATRIUM therefore predicts that procedural strategies should dominate in all three benchmarks.

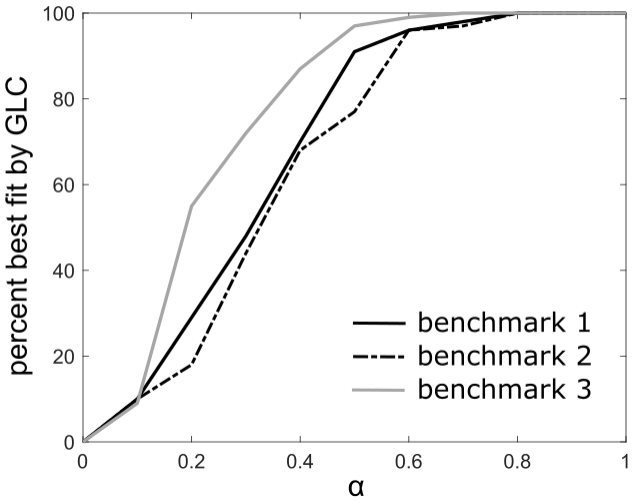

Figure 4 *The number of the 100 simulated participants whose responses were best fit by the GLC for 11 different values of* α.

6.3 Models that Switch Systems When Explicit-System Accuracy is Low

Recall that these models base the decision to switch systems solely on the performance of the explicit system. Four different versions of this general model were investigated, depending on whether switching was hard or soft, and depending on how explicit-system performance was assessed (see Table 1).

It turns out that all four of these models are easy to reject. First, consider the versions that use the COVIS explicit-system weight $\theta_{\mathrm{E}}(n)$ to monitor explicit-system performance. Specifically, either a permanent switch is made on the first trial for which $\theta_{\mathrm{E}}(n) < C_{\mathrm{E}}$, for some criterion C_{E} (i.e., model hard-E-θ), or else the procedural system is used on all trials for which $\theta_{\mathrm{E}}(n) < C_{\mathrm{E}}$ and the explicit-rule system is used on all trials for which $\theta_{\mathrm{E}}(n) \geq C_{\mathrm{E}}$ (i.e., model soft-E-θ).

Recall from section 4.1 that if $\Delta_+ = \Delta_- = \Delta$ on each trial and if Δ decreases over trials (cooling, in the parlance of neural networks), so that the weight on trial n equals $1/n$ then θ_{E} equals the proportion of trials on which the explicit system responds correctly. In models without cooling, like COVIS, in which the values of Δ do not change over trials, θ_{E} will fluctuate trial-by-trial and not converge to a single value as does proportion correct. Even so, so long as Δ_+ and Δ_- are equal or nearly equal, the fluctuations will tend to be about the proportion of trials that the explicit system responds correctly. This causes a problem for models that switch to procedural responding when explicit system accuracy drops below a criterion value of C_{E}. This is because explicit system accuracy is considerably lower in benchmark 2 (i.e., 56% correct) than in benchmark 1 (i.e., 82% correct). Therefore, for any parameter values for which Δ_+ and Δ_- are not too different, θ_{E} will tend to be significantly smaller in benchmark 2 than in benchmark 1. This means that any value of the criterion C_{E} that causes a switch to procedural responding in benchmark 1 will also cause a switch to procedural responding in benchmark 2. In other words, any such model must either predict rule-based responding in benchmarks 1 and 2 (i.e., when C_{E} is extremely

high) or else procedural responding in both tasks. Either way, one of the two benchmark tests is failed.

This problem is illustrated in Figure 5, which plots the value of the COVIS explicit system weight θ_E against trial number for benchmarks 1 and 2. These curves were generated under the assumption that $\Delta_+ = \Delta_- = .02$. In benchmark 1, θ_E fluctuates up and down around the value .82, whereas in benchmark 2 it fluctuates above and below the value .56. To successfully account for these benchmark tests requires a criterion value C_E that causes the model to switch to the procedural system in benchmark 1, but not in benchmark 2. Note that the version of the model that switches back and forth depending on whether θ_E is larger or smaller than C_E (i.e., model soft-E-θ) requires a larger value of C_E than the version that makes a single permanent switch (i.e., model hard-E-θ). The latter version will switch if and only if the minimum value of $\theta_E < C_E$, whereas the former version will respond procedurally in benchmark 1 more often than with an explicit rule when C_E is larger than its median value (i.e., so greater than 0.82). But note from Figure 5 that either value of C_E will cause the model to switch to the procedural system almost immediately in benchmark 2. In other words, because the explicit system accuracy is so much worse in benchmark 2 than in benchmark 1, explicit-system accuracy bad enough to cause a switch to a procedural strategy in benchmark 1 would also cause a switch in benchmark 2. This prediction strongly contradicts the literature since Ell and Ashby (2006) reported that 70% of participants responded in benchmark 2 as if they were using an explicit rule.

When Δ_+ and Δ_- are substantially different, then θ_E will typically be quite different from the proportion of trials on which the explicit system responds correctly. But even in these cases, it turns out that θ_E always tends to be lower in benchmark 2 than in benchmark 1. This is illustrated in Table 2, which shows mean values of θ_E at the end of learning in benchmark tests 1 and 2 for various values of Δ_+ and Δ_-. Note that when $\Delta_+ = \Delta_-$ then as expected, these values are not too different from the accuracy of the explicit system (i.e., .82 and .56 in benchmarks 1 and 2, respectively). When $\Delta_+ \neq \Delta_-$, however, the values of θ_E sometimes differ substantially from the explicit system accuracy. Even so, note that in every case, θ_E is smaller in benchmark 2 than in benchmark 1. In some cases, the differences are not large. Even so, in every case, the value of θ_E in benchmark 2 is more than 1.5 standard deviations smaller than the value in benchmark 1. As a result, all versions of this model – that is, for all parameter values for both the hard- and soft-switch versions – predict that if a switch to procedural responding occurs in benchmark test 1 then it will also occur in benchmark test 2.

Next, consider the versions of the model that switch to the procedural system when the proportion of errors in the last N_t trials is greater than β_E, where N_t and β_E are free parameters (i.e., models hard-E-N and soft-E-N). These versions are equally easy to reject. This should be expected since they differ only in how performance is assessed. As a result, they face the same problem as the models that switch systems when θ_E is low – namely, performance is worse in benchmark 2 than in benchmark 1, no matter how it is measured, so any criterion that causes a switch to procedural strategies in benchmark 1 will almost surely also cause a switch in benchmark 2. As a result, like the models that monitor θ_E, the models that switch when the proportion of errors in the last N_t trials exceeds β_E either predict rule-based responding in both benchmarks 1 and 2, or procedural responding in both tasks.

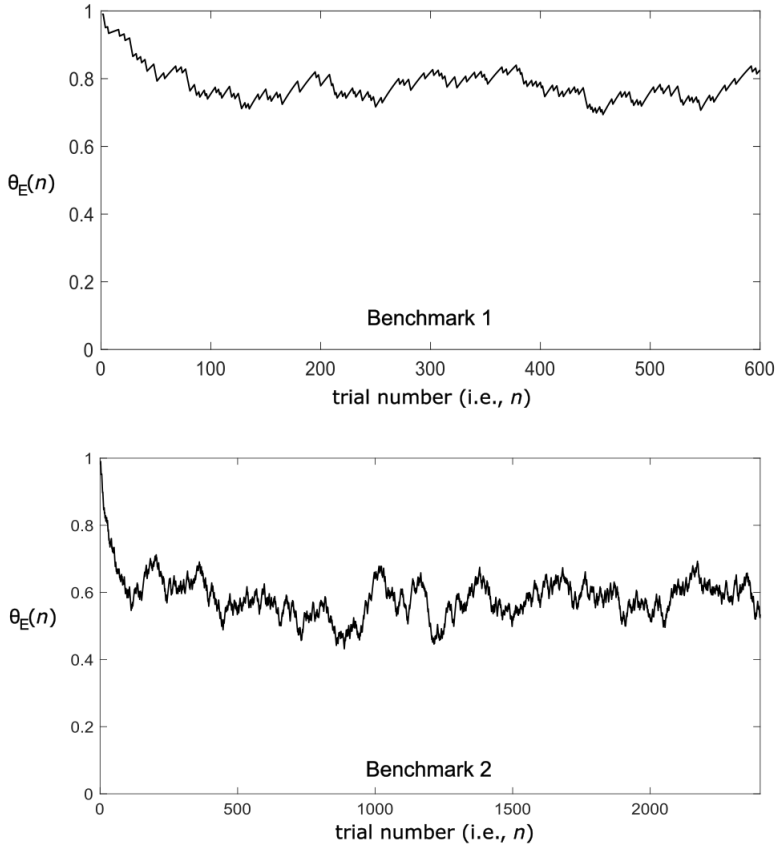

Figure 5 *The value of the COVIS explicit system weight θ_E plotted against trial number for benchmarks 1 (top panel) and 2 (bottom panel). Both plots were generated with $\Delta_+ = \Delta_- = .02$.*

Table 2 *Mean values of θ_E at the end of learning for various values of Δ_+ and Δ_-. B1 = benchmark test 1; B2 = benchmark test 2.*

		Δ_+					
		0.001		**0.101**		**0.201**	
		B1	**B2**	**B1**	**B2**	**B1**	**B2**
Δ_-	**0.001**	.886	.604	.997	.992	.998	.996
	0.101	.030	.013	.765	.569	.863	.724
	0.201	.015	.006	.609	.396	.757	.562

As an illustration of the problems these models face, consider first model hard-E-N. As an example of the extreme difficulty this model faces, consider benchmark test 2 when $N_t = 20$. In the simulations I ran, when $\beta_E \leq .70$ the model always eventually switched to the procedural system and therefore the last 100 responses were always best fit by the GLC. Even when $\beta_E = .75$, the GLC fit better than the 1D model for 85 of the 100 simulated participants. The smallest value of β_E for which the 1D model provided the better fit to the majority of participants was $\beta_E = .80$, but even then, the GLC still fit better for 43 of the 100 participants. With $\beta_E = .80$ and $N_t = 20$, a switch to procedural responding only occurs when there are 3 or fewer correct responses on the last 20 trials. The (binomial) probability that this occurs with chance responding is approximately $p = .001$. But the Ell and Ashby (2006) study that inspired benchmark 2 trained their participants for 2,400 trials each. This is enough trials to make it reasonably likely that an event with probability $p = .001$ will occur at least once. In fact, of the 100 simulated participants, 38 switched systems when $\beta_E = .80$. A value of $\beta_E \geq .75$ is unreasonable though. For example, in benchmark 1, the best one-dimensional rule achieves an accuracy of 82% correct. So with $\beta_E \geq .75$, this model would never switch to a procedural strategy in benchmark 1 because when $p = .82$, a 20-trial run with 4 or fewer correct responses is essentially impossible.

The version that makes multiple switches because it uses the explicit system on all trials for which the proportion of errors in the preceding N_t trials is greater than β_E (i.e., model soft-E-N) is equally easy to reject. For example, with $N_t = 20$, the last 100 responses of the model are best fit by the GLC for 54% of the simulated participants in benchmark 1 when $\beta_E = .20$ and for 86% when $\beta_E = .15$. So to pass benchmark test 1, β_E must be less than or equal to .15. However, with $\beta_E = .15$, the GLC fits best for all simulated participants in benchmark 2 (i.e., with $N_t = 20$). So this version of the model badly fails benchmark 2. To pass benchmark 2, β_E must be set to at least $\beta_E = .40$. In this case, the final 100 responses of 58% of simulated participants are best fit by the 1D model. However, with $\beta_E = .40$, the responses of 97% of benchmark 1 participants are also best fit by the 1D model. So this version of the model badly fails benchmark 1.

6.4 Models that Monitor Procedural-System Accuracy

Recall that these models always begin with the explicit system, and switch to the procedural system only if the performance of the explicit-system is poor *and* procedural-system performance is good. I investigated the analogous four versions of this model as in the previous subsection (i.e., see Table 1). Specifically, the four versions differed according to whether the COVIS system weights or the number of errors in the preceding N_t trials were used to assess performance and in whether system switching was hard or soft.

For each of the four versions of this model, it turned out to be easy to find parameter values for which the model passed all three benchmark tests. Table 3 describes the performance of successful examples of each of the four models. Specifically, for each model, Table 3 lists the percentage of simulated participants in each benchmark whose responses were best fit by the GLC and 1D classifier, along with the mean and standard deviation of the BIC scores of the best-fitting versions of each model. To assess the robustness of these predictions, I systematically explored how sensitive they are to changes in each of the model's parameters. This process allowed a crude approximation of the volume of parameter space over which each model passed all three benchmark tests.

Table 3 *Benchmark test results for the models that monitor procedural-system accuracy. (soft-EP-θ = soft-switch model that uses COVIS system weights to monitor accuracy: C_E = .8, C_P = .75, Δ = .02; hard-EP-θ = hard-switch model that uses COVIS system weights to monitor accuracy: $C_E = .7, C_P = .77, \Delta = .02$; soft-EP-N = soft-switch model that uses number of recent errors to monitor accuracy: N_t = 30; β_E = .17, β_P = .20; hard-EP-N = hard-switch model that uses number of recent errors to monitor accuracy: N_t = 30; $\beta_E = .33$, $\beta_P = .13$).*

Model		Benchmark 1 (II Low Overlap)		Benchmark 2 (II High Overlap)		Benchmark 3 (Hybrid)	
		GLC	1D	GLC	1D	GLC	1D
soft-EP-*θ*	win %	98	2	2	98	11	89
	mean BIC	76.6	100.8	96.0	92.5	47.0	44.1
	BIC std	12.6	13.1	10.9	10.8	13.1	14.7
hard-EP-*θ*	win %	89	11	14	86	3	97
	mean BIC	67.3	98.5	100.0	99.7	43.7	40.3
	BIC std	12.2	14.4	13.3	19.2	8.5	9.0
soft-EP-*N*	win %	90	10	3	97	25	75
	mean BIC	80.5	97.3	96.5	93.0	54.5	53.6
	BIC std	13.7	13.3	11.5	11.4	15.8	18.8
hard-EP-*N*	win %	98	2	17	83	29	71
	mean BIC	66.1	102.0	100.9	101.4	45.9	54.5
	BIC std	13.0	12.2	14.1	20.9	11.5	28.3

Model soft-EP-θ

Model soft-EP-θ uses the procedural system on all trials for which $\theta_E < C_E$ and $\theta_P > C_P$ and uses the explicit system on all trials when one or both criteria fail. The model has three free parameters: C_E, C_P, and Δ.

The predictions of model soft-EP-θ are quite robust. First, all three benchmark tests are passed for any value of Δ between .004 and .201. For example, even with a ten-fold increase in the value of Δ that was used to generate Table 3 (i.e., from .02 to .201), the model still passes all benchmarks, with the GLC fitting best for 80% of simulated participants in benchmark 1, and the 1D model fitting best in benchmarks 2 and 3 for 83% and 61% of participants, respectively. On the other hand, if Δ is too small – specifically, if $\Delta \leq .003$ – then learning is so slow that θ_P never grows large enough in benchmark test 1 to cause a switch to procedural responding. For example, when $\Delta = .003$, the GLC fits best for only 7% of simulated participants in benchmark test 1.

The model also passed all benchmarks for a broad range of the parameter C_P – specifically for values of C_P in the range [.65, .85]. For example, when $C_P = .65$ (i.e., and with $C_E = .8$, and $\Delta = .02$), the GLC fit better than the 1D model for 95%, 23%, and 8% of simulated participants in benchmarks 1, 2, and 3, respectively. When $C_P = .85$, these percentages were 89%, 3%, and 7%. I made no attempt to find the exact upper and lower values of C_P that allow the model to pass all benchmarks, but it is important to note that

these values exist. For example, when $C_P = .55$, the GLC fit best in benchmark 2 for 92% of simulated participants – a value that is strongly inconsistent with the empirical data. So when $C_P = .55$, the model switches to procedural responding even though procedural accuracy is poor, whereas the data suggest that humans will persist with rule-based strategies under these conditions. When C_P is too large, the problem is that the model never switches to procedural responding, even in benchmark test 1. For example, with $C_P = .95$, the GLC fit best for only 2% of simulated participants in benchmark 1.

Similar, but narrower, upper and lower limits exist for the parameter C_E. The model passes all three benchmarks for values of C_E in the range [.77, .83]. The model fails at least one benchmark for values outside of this range. If C_E is too low then the model never switches to procedural responding – rather it remains satisfied with poor rule-based performance. For example, with $C_E = .75$, the GLC fit best for only 49% of simulated participants in benchmark 1 – a value that is much less that reported in typical empirical applications. When C_E is too high, then the model gives up on rule-based responding more easily than human participants. For example, with $C_E = .85$, the GLC fit best for 63% of simulated participants in benchmark 3 – a value that is considerably higher than reported by Ashby and Crossley (2010).

Model hard-EP-θ

Model hard-EP-θ is the same as model soft-EP-θ, except it makes a single hard switch to the procedural system on the first trial that the COVIS explicit system weight θ_E drops below some value C_E and at the same time, θ_P is greater than some value C_P. The model has the same three free parameters as model soft-EP-θ; namely, C_E, C_P, and Δ.

This model passed the three benchmarks over a more limited region of its parameter space than model soft-EP-θ. As with the soft version, if Δ is too small, the learning is too slow to cause a shift to procedural responding in benchmark test 1. For example, when $\Delta = .015$, the responses of all simulated participants in benchmark 1 were best fit by the GLC, whereas when $\Delta = .01$, 80% of simulated participants persisted with a rule. Unlike the soft version, however, model hard-EP-θ badly failed the tests for larger values of Δ. For example, when $\Delta = .04$, the model switched to procedural responding in benchmark 2 for 97% of simulated participants. In fact, the model passed all three benchmarks only for the narrow range $.015 \leq \Delta \leq .025$ (i.e., when $C_E = .7$ and $C_P = .77$). The problem is that a large learning rate Δ causes large fluctuations in θ_E and θ_P. This has enormous consequences for the model because a single uncharacteristic deviation in benchmark 2, for example, can cause the model to switch permanently to procedural responding. The same fluctuations also cause a switch to procedural responding in model soft-EP-θ, but because switching is soft the model will switch back to rule-based responding when the fluctuation is reversed. A few trials of procedural responding during the final 100 trials do not cause the GLC to fit better than the 1D model.

The ranges on C_E and C_P over which the model passed all benchmarks were also narrow. For C_E this range was approximately $.69 \leq C_E \leq .75$. For values below this range, the model was too satisfied with poor performance in benchmark 1, and as a result, never switched to procedural responding (at least, not nearly as often as humans). For values above this range, the model gives up on rules more easily than humans, and as a result, for example, switches to procedural strategies in benchmark 3 at a much higher rate

than human participants. For C_{P}, the model passes all three benchmarks over the range $.75 \leq C_{\mathrm{P}} \leq .87$. Values below this range cause the model to fail benchmark 2, and values above the range cause it to fail benchmark 1.

Model soft-EP-N

Model soft-EP-N soft switches to the procedural system on all trials for which the proportion of explicit system errors during the preceding N_t trials is greater than β_{E} *and* the proportion of procedural system errors during these same trials is less than β_{P}. It uses the explicit system on all trials when one or both criteria fail. The model has three free parameters: N_t, β_{E}, and β_{P}.

The predictions of the model are robust over changes in the parameter N_t. When $\beta_{\mathrm{E}} = .17$ and $\beta_{\mathrm{P}} = .20$ (as in Table 3), the model passes all three benchmarks for values of N_t in the range [15, 100]. For values of $N_t < 15$, the model switches so frequently that the last 100 simulated trials of benchmark test 1 include more rule-based responding than the empirical data. For example, when $N_t = 6$, the 1D model provided the better fit to 54% of simulated participants. At the other extreme, I did not investigate predictions for values of $N_t > 100$. When $N_t = 100$, the model easily passes all three benchmarks, so the true upper limit is likely considerably higher than this.

In contrast, the model is quite sensitive to changes in both β_{E}, and β_{P}. First, any drop in β_{E} causes the model to fail benchmark 3 because when $\beta_{\mathrm{E}} < .17$, the model gives up too easily on rules.[6] At the other extreme, the model passes all benchmarks when $\beta_{\mathrm{E}} = .20$, but larger values cause the model to fail benchmark 1. Specifically, with a large value of β_{E}, the model is more satisfied with poor explicit system performance than human participants. The range on β_{P} over which the model passes all three benchmarks is approximately [.15, 25]. When $\beta_{\mathrm{P}} < .15$, the criterion on procedural-system performance is too high and as a result, the model fails benchmark 1 – that it, it persists with rules more frequently than human participants. When $\beta_{\mathrm{P}} > .25$, the criterion on procedural-system performance is too lax and it now switches to procedural responding more often than human participants. For example, it now fails benchmark test 2.

Model hard-EP-N

Finally, model hard-EP-N is identical to soft-EP-N except it makes a single hard switch to the procedural system on the first trial for which the proportion of explicit-system errors during the preceding N_t trials exceeds β_{E} *and* the proportion of procedural-system errors is less than β_{P}. This model also has three free parameters: N_t, β_{E}, and β_{P}.

Model hard-EP-N is more sensitive to sample size N_t than model soft-EP-N. Specifically, it passes all three benchmark tests for values of N_t in the range [28, 50], which is considerably narrower than the range over which model soft-EP-N passes all benchmarks (i.e., [15, 100]). Both models appear to be about equally sensitive to changes in the parameters β_{E} and β_{P}. Specifically, the range on values of β_{E} over which model hard-EP-N passes all benchmarks is approximately [.30, .36], whereas the range on β_{P} is approximately [.09, .13].

[6] More specifically, this statement assumes a drop large enough to cause the criterion on the number of explicit-system errors during the last $N_t = 30$ trials to change by at least one.

Reconsidering benchmark test 3

Recall that when I simulated responses in benchmark test 3, I assumed that the procedural system of all models used a suboptimal linear bound with slope 1 and intercept 0. The models that were rejected failed benchmark 3 because they incorrectly predicted that procedural responding should dominate. For example, for all 9 of the COVIS parameter combinations that I investigated, the model failed benchmark 3 because it predicted that procedural responding should dominate the final 100 responses. All models privilege the more accurate of the two systems, so my choice to assume a suboptimal decision bound when simulating benchmark 3 responses was made to bias the tests in the model's favor. The fact that they failed to predict rule-based responding, despite this bias, strengthens our conclusions that the models are fundamentally flawed.

On the other hand, for the models considered in this subsection that passed all benchmarks, the bias seems inappropriate. Specifically, how can we rule out the possibility that the four models identified in Table 3 passed benchmark 3 only because they were assumed to use a suboptimal bound? To investigate this question, I reran all four models through benchmark test 3 with exactly the same parameter values listed in Table 3, except under the assumption that the model's procedural system used the optimal (nonlinear) decision bound on every trial. In every case, the models again easily passed benchmark 3 – that is, they all continued to predict that most participants would persist with a rule throughout the last 100 trials. Why does it make no difference whether the models' procedural system uses a suboptimal linear bound that achieves an accuracy of 90% or an optimal nonlinear bound that achieves 100%? The key is that the best one-dimensional rule performs so well in benchmark 3 (i.e., 90% correct) that the models never become dissatisfied with their explicit system performance. As a result, they never switch systems, despite the marginal accuracy gains that are possible.

7. DISCUSSION

7.1 Models Rejected by These Results

COVIS assumes that system switching is a routine and common occurrence and that soft, back-and-forth, trial-by-trial system switching occurs frequently throughout category learning. In fact, the model predicts that trial-by-trial switching is common even after performance asymptotes – that is, after learning is essentially complete. COVIS easily passes benchmark test 1, but badly fails tests 2 and 3 – primarily because it is overly optimistic about performance in these tasks. The COVIS system weights tend to oscillate around the true probability that the associated system responds correctly. As a result, COVIS tends to privilege the more accurate of the two systems. All three benchmark tests use an II category structure. Under the training conditions that these tests assume (e.g., immediate feedback), the COVIS procedural system is therefore more accurate than the COVIS explicit system, and as a result, COVIS predicts a significant contribution of procedural responding in all three tests. This prediction is supported in test 1 and strongly violated in tests 2 and 3. In addition, COVIS also fails to account for the extreme difficulty participants have in switching between declarative and procedural strategies in the few tasks that were designed to study this ability (Ashby & Crossley, 2010; Crossley et al., 2018; Erickson, 2008).

ATRIUM assumes the outputs of explicit-rule and procedural/exemplar-similarity systems are blended and therefore that there is no system switching of any kind. Instead, both systems contribute to the categorization response on every trial, with the magnitude of their contributions determined by the mixing parameter α (i.e., see Equation 6). Figure 4 shows that there is no single value of α that allows the model to pass the three benchmark tests. ATRIUM assumes that the asymptotic value of α should reflect the categorization success of the procedural/exemplar-similarity system, relative to the explicit-rule system. In benchmark tests 1 and 2, the procedural/exemplar-similarity system significantly outperforms the best rule, and so ATRIUM predicts that α should be greater than .5 in these tasks. In benchmark 3, the two systems are equally accurate, so α should have a value near .5. Given such values of α, ATRIUM predicts that procedural strategies should dominate in all three tests (i.e., see Figure 4). As a result, like COVIS, ATRIUM fails benchmark tests 2 and 3.

There is also other strong evidence against any model that assumes blending. In particular, many studies have compared performance in RB and II tasks across a variety of different training and testing conditions. In many cases, the categories used in the two tasks were matched on most category-separation statistics. The studies varied the methods and order of stimulus presentation, the nature and timing of feedback, the quality of the training conditions, and the generalizability of the knowledge acquired during training. This line of research has produced scores of articles that have reported more than 30 qualitatively different dissociations between training and performance in RB and II tasks (for a review of many of these, see Ashby & Valentin, 2017). In other words, the experimental manipulation had a significant effect on performance in one task, but not the other. In some studies, the significant effect was on the RB task and in others, it was on the II task. These results were all predicted a priori by COVIS, which assumes that the primary effect of these manipulations was on the one system that dominates performance in the affected task. In contrast, if both systems contribute to the response on every trial in all tasks, as predicted by blending (and ATRIUM), then one would expect more graded results in which the manipulation affected both tasks, albeit one more than the other. So any model that assumes blending is challenged to account for the many reported empirical dissociations between learning and performance in matched RB and II tasks.

All models in which the decision to switch systems depends only on the accuracy of the explicit system fail the benchmark tests because explicit system accuracy is lower in benchmark 2 than in benchmark 1. As a result, if a switch to procedural responding occurs when explicit system accuracy drops below some criterial value then all such models must fail either benchmark 1 or benchmark 2. One possibility is that the criterion is set to such a high value that a switch to procedural responding doesn't occur in either benchmark. In this case, the model would fail benchmark 1 and pass benchmark 2. The second possibility is that the criterion is set low enough that a switch to procedural responding occurs in benchmark 1. But since explicit-system accuracy is lower in benchmark 2 than benchmark 1, this means that procedural strategies should also dominate in benchmark 2 – in other words, that the model would fail benchmark test 2.

One revision that might rescue models in which the decision to switch depends only on explicit-system accuracy is to allow different thresholds for switching in the different benchmark tests. The idea here is that the participant becomes aware that optimal ac-

curacy is considerably higher in benchmark 1 than in benchmark 2 – perhaps because of experimenter instruction. Suppose further that the participant adopts a satisficing strategy that uses the explicit system so long as performance remains acceptably close to this optimal value. For example, such a participant might decide that 56% correct is acceptable in benchmark 2, whereas 82% is unacceptable in benchmark 1. This revision to the model provides a plausible rationale for why the switching threshold might be set to a significantly higher value in benchmark 1 than in benchmark 2.

Although this more general class of models could account for the three benchmark tests that were the focus of this article, there are other, well-replicated results that make it easy to reject all models of this type. In particular, many studies have reported that certain experimental manipulations impair performance in II tasks, like the one described in benchmark test 1, more than in the RB task that is matched on most category-separation statistics (i.e., in which the RB categories are created by rotating the II categories in stimulus space to have a vertical or horizontal orientation). For example, II learning is impaired with feedback delays as short as 2.5 s, whereas RB learning is unaffected by feedback delays as long as 10 s (Maddox et al., 2003; Maddox & Ing, 2005). Similarly, II learning is impaired if the feedback is batch delivered after every six trials, or if the feedback about the accuracy of a response is given immediately after the response on the ensuing trial, whereas RB learning is unaffected by either of these manipulations (J. D. Smith et al., 2014, 2018). Critically, however, in all of these studies, a strategy analysis revealed that virtually all participants used an explicit-rule strategy in the II condition. In other words, in benchmark test 1, almost all participants switch from an explicit rule to a procedural strategy when procedural accuracy is high, but they fail to switch with essentially the same category structures when experimental conditions impair or abolish procedural learning. The inescapable conclusion seems to be that we can reject any model that does not monitor procedural system accuracy.

7.2 Models That Pass All Benchmark Tests

In sharp contrast to all other models, models that switch systems when explicit-system accuracy drops below a criterion value but only if the procedural system has been performing reasonably well, easily pass all three benchmark tests. In benchmark test 1, a switch is made because procedural-system accuracy is nearly perfect, whereas explicit system accuracy is only 82% correct. In benchmark 2, no switch is made because although the explicit system performs poorly, so does the procedural system. Finally, there is also no switch in benchmark 3 – this time because explicit system accuracy is satisfactorily high. These results are quite robust, in the sense that they do not depend on whether switching is hard or soft or on exactly how the performance of each system is monitored. Table 3 shows that four different versions of this model easily pass all three tests.

Although the results reported in Table 3 suggest considerable diversity in the class of models that pass the benchmark tests, any model that switches to procedural responding only when procedural accuracy is reasonably high must make several strong assumptions. First, all such models must assume that the procedural system learns during trials when the explicit system controls responding. Without such background learning, procedural accuracy would remain at chance and as a result, no system switch would ever occur. Second, the supervisory network that determines which system will control responding must

have an estimate of procedural system accuracy, even on trials when the explicit system is controlling behavior. How realistic are these assumptions?

The evidence that procedural learning occurs while responding is under declarative control is reasonably strong. The strongest evidence comes from a study that directly tested this hypothesis. Crossley and Ashby (2015) trained two groups of participants on stimuli drawn from II categories, like those shown in Figure 1, for 800 trials each. For the control group, all 800 trials consisted of standard II categorization training. For the experimental group, the first 600 trials were divided into two one-dimensional RB tasks in which all stimuli were selected from the II categories. In one RB task, dimension 1 was relevant and category A included all category A stimuli from the II categories that were to the left of the RB bound, whereas the category B stimuli included all category B stimuli from the II categories that were to the right of the 1D bound. II stimuli that did not meet these criteria were excluded from these trials. The second RB task followed a similar logic, except with dimension 2 as the relevant dimension. Now categories A and B were created by sampling stimuli from the II categories that were above and below the RB optimal bound. Given this design, perfect accuracy was possible in both RB tasks. Following this training, participants had correctly categorized all stimuli in the II categories, but by using an explicit rule rather than a procedural strategy. Next, during the final 200 trials, participants transferred to the full II categories. Results showed that there was no difference between the two groups on the final 200 II trials. In other words, even though the experimental participants had no prior II training and the control group had been practicing on the II categories for 600 trials, the experimental participants were just as good as the control participants during these last 200 II trials.

To verify that the good performance of the experimental group was because of procedural learning that occurred during the first 600 RB trials, Crossley and Ashby (2015) ran a follow-up experiment that was essentially identical, except feedback was delayed for 2.5 sec during the RB trials (but not during the II trials). Prior research had shown that feedback delays of this duration impair procedural learning, but not explicit, rule learning. The results showed no transfer of knowledge from the RB training to the II trials. These results would be impossible if procedural learning did not occur in the background during declarative control of responding.

Other results support this conclusion. For example, Kalra et al. (2025) provided participants with an explicit, verbalizable rule that perfectly categorized two sets of stimuli. Even though instructed to use this rule on every trial, results showed that participants also used another dimension, not mentioned in the rule they were taught, that also conveyed some diagnostic information that was assumed to require procedural learning to acquire. There is also other, more indirect evidence that procedural learning occurs in the background. For example, consider benchmark test 1. Considerable evidence suggests people begin with simple, explicit rules in this task and only switch to procedural strategies later in training. If there was no procedural learning during these early explicit-rule trials, then when the switch to procedural control is made, accuracy would immediately drop to chance before incrementally increasing again as procedural learning progresses. In the case of the benchmark 1 categories, this drop would be precipitous since a 1D rule can achieve an accuracy of 82% correct. Despite scores of published studies that used II categories like those shown in Figure 1, none have reported learning curves that display a conspicuous drop of

this type.

Unfortunately, I know of no studies that examined the question of whether the supervisory network that determines which system controls responding estimates procedural system accuracy on trials when the explicit system controls behavior. It is not even clear how to design an experiment that directly examines this issue. The best evidence in support of this possibility probably comes from the results described above. Specifically, it is not clear how any other hypothesis could account for the following three results: most participants 1) switch to a procedural strategy with II categories like those of benchmark test 1 when feedback is immediate; but 2) persist with explicit-rule strategies with these same categories when procedural learning is impaired by a feedback delay; and 3) persist with explicit-rule strategies in benchmark 2 (i.e., when procedural accuracy is low). Clearly, more research is needed on this question.

Table 3 indicates that four different models that monitor the accuracy of both the explicit system and the procedural system pass all three benchmark tests. As a result, the present analyses can neither favor nor disfavor any of these models. Even so, there are some differences among the models worth noting. One is that the range on the parameters that allowed the hard-switch models to pass the benchmarks was narrower than the range on the parameters of the soft-switch models. This was true for all parameters, and in some cases, a hard-switch model passed the benchmarks only for an extremely narrow parameter range. For example, model hard-EP-N passed the benchmarks only when $.09 \leq \beta_{\text{P}} \leq .13$. So one advantage of the soft-switch models is that their predictions are more robust than the predictions of the hard-switch models.

Decision-bound modeling of the type used in this article assumes stationary data – that is, it assumes that the participant used the same decision strategy on every trial of the to-be-modeled data (e.g., Ashby, 2026). This is the reason the models were fit only to the last 100 trials of simulated responses. Hélie et al. (2017) developed an iterative version of decision-bound modeling that estimates all response strategies ever used by a participant, along with the trial number on which each system-switch occurs. They then used this model to re-analyze the data of Ell and Ashby (2006), which included the experiment that is the basis of benchmark test 3. They concluded that participants switched strategies infrequently, but more often than assumed by the hard-switch models. Their results are therefore most consistent with versions of model soft-EP-θ in which Δ is small and versions of model soft-EP-N in which N_t is large. This is because a small Δ and a large N_t reduce the number of strategy switches.

7.3 Toward a Possible Neural Model

The evidence is good that explicit-rule learning is mediated by a broad neural network centered in the prefrontal cortex (PFC), whereas the most important neural area in procedural learning is the striatum (e.g., see Ashby, 2026). The PFC projects to the striatum (e.g., head of the caudate nucleus), but not to regions of the striatum thought to mediate procedural learning (e.g., body and tail of the caudate nucleus, posterior putamen). Furthermore, no parts of the striatum project directly to the PFC. As a result, the two systems almost surely do not directly interact. Instead, the neuroanatomical evidence suggests that hierarchical control is mediated by some separate supervisory network. One possible model of that network is illustrated in Figure 6.

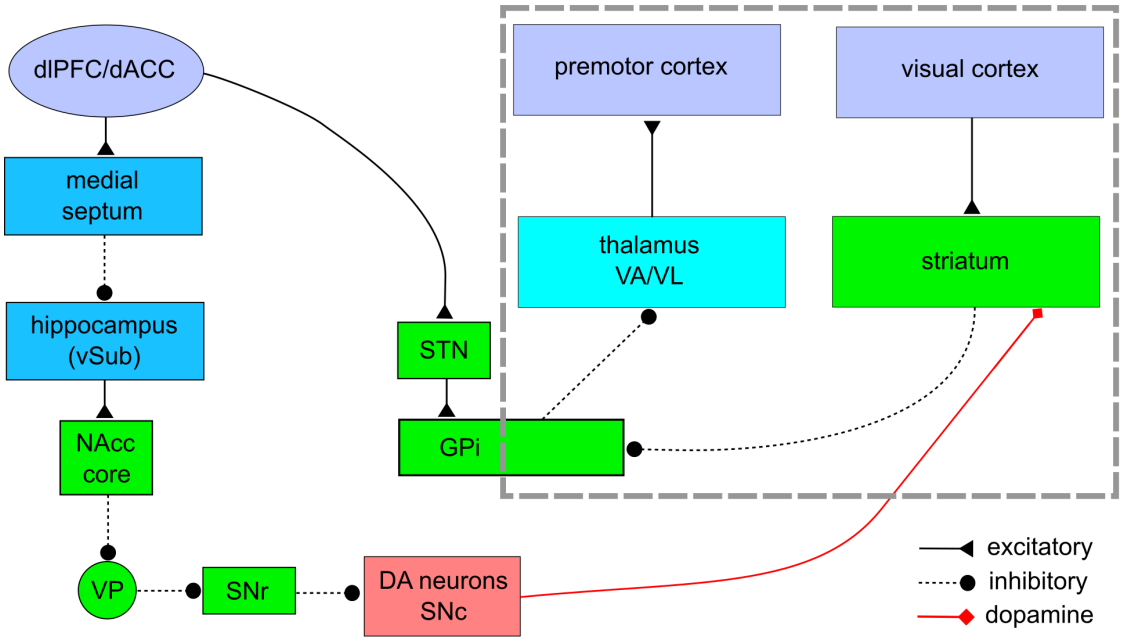

Figure 6 *A neural model of system switching. The broken gray rectangle identifies regions that represent a simplified version of the procedural-learning system. dlPFC = dorsolateral prefrontal cortex; dACC = dorsal anterior cingulate; vSub = ventral subiculum; NAcc = nucleus accumbens; VP = ventral pallidum; SNr = substantia nigra pars reticulata; DA = dopamine; SNc = substantia nigra pars compacta; STN = subthalamic nucleus; GPi = internal segment of the globus pallidus; VA = ventral anterior nucleus; VL = ventral lateral nucleus.*

There are three separate subnetworks illustrated in Figure 6. The network inside the broken gray rectangle is a simplified model of the procedural-learning system. Current models are more complex than this. For example, Cantwell et al. (2015) proposed and presented evidence that procedural learning is mediated by double cortical-striatal-cortical loops, rather than the single loop shown in Figure 6.

The second network illustrated in Figure 6 is the hyperdirect pathway that begins with direct excitatory projections from PFC to the subthalamic nucleus, which then sends excitatory projections to the internal segment of the globus pallidus (GPi). The GPi, which is a major output structure of the basal ganglia, serves as a gate between the striatum, where the synaptic plasticity that mediates procedural learning is thought to occur, and motor areas of cortex that execute motor actions mediated by the procedural system. The GPi inhibits downstream targets (in thalamus), and procedural responses selected in the striatum are enacted when the striatum selectively inhibits GPi activity, thereby disinhibiting the selected motor action. Ashby and Crossley (2010) proposed that switching from an explicit rule to procedural responding is mediated by the hyperdirect pathway. The idea is that while the explicit system is in control, the PFC excites the subthalamic nucleus, which then excites the GPi, and this extra excitatory input to the GPi offsets any inhibitory input from the striatum, thereby making it more difficult for striatal activity to affect motor areas of cortex.

In this way, the hyperdirect pathway could permit (by reducing subthalamic activity), or prevent (by increasing subthalamic activity) a procedural response selected in the striatum from reaching motor cortex. The primary evidence implicating the hyperdirect pathway in system switching comes from studies using the stop-signal task, in which participants must respond quickly to a cue but on some trials a second cue signals that they must inhibit this response. The subthalamic nucleus is known to play a key role on trials when the cued motor response must be stopped (Aron & Poldrack, 2006; Aron et al., 2007; Pasquereau & Turner, 2017). A popular model is that the second stop cue causes PFC to excite the subthalamic nucleus thereby canceling the motor command being sent through the striatum.

The third network illustrated in Figure 6 describes a pathway from the PFC to the dopamine (DA) neurons in the substantia nigra pars compacta (SNc) that enervate the striatal regions that mediate procedural learning. To my knowledge, this is the first article to propose that this subnetwork participates in supervisory control of rule-based versus procedural responding. This pathway was identified by Bortz and Grace (2018), who showed that activation of the medial septum in rats decreased the number of spontaneously active DA neurons in the SNc. Furthermore, Bortz et al. (2023) showed that activation of this pathway improved the ability of rats to switch from an operant strategy they had been using to an older strategy that had not been used in some time. Activation of this medial septum pathway will have similar effects on striatal output as activation of the hyperdirect pathway. This is because DA increases the signal-to-noise ratio of cortical input to the striatum by potentiating the glutamate response through NMDA receptors and depressing the glutamate response through non-NMDA receptors (Ashby & Casale, 2003). As a result, inhibiting the SNc DA neurons will reduce the magnitude of striatal output to the GPi, thereby reducing the likelihood that a striatal-mediated behavior is executed. For example, this is exactly how the death of DA neurons in the SNc causes bradykinesia in Parkinson's disease.

Bortz and Grace (2018) identified another feature of the medial-septum pathway (not shown in Figure 6 to keep the figure as simple as possible), which makes it an even more attractive model of system switching. Specifically, they showed that activation of the medial septum not only reduces the number of tonically firing DA neurons in the SNc, but it simultaneously has the opposite effect on DA neurons in the ventral tegmental area (VTA) – that is, it increases the number that are tonically active.[7] The VTA sends a prominent DA projection to frontal cortex and to virtually all neural areas that have been implicated in rule learning. Increasing the number of tonically active DA neurons in the VTA will therefore increase DA levels in the neural network that defines the explicit system. This is important because there is substantial evidence that all forms of executive function are facilitated when DA levels in frontal cortex increase modestly (for reviews, see, e.g., Ashby et al., 2015; Ott & Nieder, 2019). As a result, activating the medial-septum pathway should simultaneously inhibit activity in the procedural system and enhance processing in the explicit system.

Simultaneously activating the hyperdirect and medial-septum pathways shown in Figure 6 would powerfully inhibit the ability of striatal activation to reach premotor or

[7]The pathway from the medial septum to the VTA is similar to the one shown in Figure 6, except the projection from medial septum to vSub is cholinergic, rather than GABAergic, and the projection from vSub to NAcc is to the NAcc shell, rather than to the NAcc core. The VP sends GABAergic projections to both the SNr (as in Figure 6) and to the VTA.

motor cortex and simultaneously enhance the efficacy of the explicit system. As such, the simultaneous activation of both pathways would provide an especially effective method of switching from procedural to executive control of responding. On the other hand, it is unlikely that the medial-septum pathway is active during long stretches of executive control. Inhibiting the SNc DA neurons during executive control would highly impair or prevent any background procedural learning, whereas as previously mentioned, the evidence is good that the procedural system learns in the background on trials that the rule system controls responding (Crossley & Ashby, 2015). So the most likely version of this model predicts that the medial septum pathway is activated on trials when a switch is made from procedural to executive control, whereas the hyperdirect pathway is continually active during periods of executive control.

This version of the Figure 6 model has some attractive features that are consistent with the results reported here. First, the model predicts that when the explicit system is in control of responding, the hyperdirect pathway inhibits procedural responding at a site that is downstream from the cortical-striatal synapses that mediate learning in the COVIS model of the procedural system. This allows for the possibility that learning at these synapses can proceed in the background during rule-based responding. How this learning occurs though, is not clear. As noted by Paul and Ashby (2013), the problem is that when the explicit system controls responding, the obtained feedback is dependent on activity within the explicit system, but not on activity within the procedural system. As a result, any change in synaptic strengths within the procedural system is as likely to be rewarded as any other change. Paul and Ashby (2013) also showed that the procedural system could learn from the explicit system if it somehow received information about the explicit system's response on each trial. To see how this might work, consider benchmark test 1. A COVIS model of the procedural system for this task would include two striatal units – one associated with each response alternative. Denote the inputs to these two units on trial n by $S_{\mathrm{A}}(n)$ and $S_{\mathrm{B}}(n)$. Now suppose that the explicit system controls responding on trial n and responds A. Then Paul and Ashby (2013) investigated predictions of a model in which the input to striatal unit B is unchanged on this trial, whereas the input to striatal unit A is changed to $S_{\mathrm{A}}(n) + H_{\mathrm{E}}(n)$, where $H_{\mathrm{E}}(n)$ is the response confidence of the explicit system as defined in Equation 1. Paul and Ashby (2013) showed that when the procedural system receives an efferent copy of the explicit system output in this way, it learns to respond as if it was using the explicit-system decision bound on trials when the explicit system controls responding.

How could the procedural system receive an efferent copy of the explicit system response? A possible solution to this problem is suggested by the neuroanatomy of the basal ganglia, which is characterized by a series of closed (as well as open) cortical-striatial-cortical loops. Closed loops of this type are prevalent in premotor and motor cortex and as a result, we know that activation in the premotor units that initiate the motor command selected by the explicit system is propagated to the striatum, and more specifically to the putamen (Seger, 2008). In the double-loop model of the procedural system proposed by Cantwell et al. (2015), the second loop projects from the putamen to premotor cortex. As a result, the procedural system will therefore receive an efferent copy of the explicit system response so long as there is overlap in these two regions of putamen.

A second attractive feature of this model is that it is unidirectional. For example, because the hyperdirect and medial septum pathways both begin with projections from the

PFC, they could be viewed as part of the explicit system. If so, then the model predicts that the explicit system is always in control. The procedural system controls responding only if allowed by the explicit system. The explicit system can control responding anytime it wants, whereas the procedural system has no way to wrest control from the explicit system. This agrees with widespread evidence that volitional control of behavior is mediated primarily by PFC (e.g., Badre & Nee, 2018; Friedman & Robbins, 2022).

Results reported in this article suggest the supervisory network monitors procedural system performance and switches from explicit-rule learning to procedural learning only if explicit performance is poor *and* procedural performance is good. How could the Figure 6 model monitor procedural accuracy? This model assumes that the decision to switch is mediated in PFC. As described above, considerable evidence suggests that the key site of procedural learning is in the striatum. This is problematic because according to the Crick-Koch hypothesis (Crick & Koch, 1998), for example, the PFC does not have direct access to striatal activity because the striatum does not project directly to PFC. If not, then how could the hyperdirect pathway model monitor procedural system accuracy? There are a number of possibilities, but in the absence of more empirical constraints, these are all speculative, and as a result will be described only briefly here. One possibility is that the supervisory network could pass control to the procedural system for a trial or two and then examine the difference in activation produced by the procedural system at the competing motor units in premotor or motor cortex. If the procedural system has been learning in the background, then this difference should be substantial (even if the response is incorrect), whereas in the absence of procedural learning, the difference should be small. A second possibility is that the supervisory network monitors cortical activity elicited by the procedural system that is upstream from the site at which the hyperdirect pathway blocks procedural-system access to motor areas of cortex. For example, Cantwell et al. (2015) proposed that procedural learning is mediated by double cortical-striatal-cortical loops. This model assumes that procedural learning is mediated by synaptic plasticity at cortical-striatal synapses on both loops, and that only the second of these loops terminates in premotor cortex. The cortical terminus of the first loop was tentatively identified as the pre-supplementary motor area[8] (preSMA). The second loop then projects from preSMA to the striatum (i.e., the putamen) then to premotor cortex via the GPi and thalamus (i.e., via its ventral-lateral nucleus). The hyperdirect pathway model would block the second of these loops from cortex, but not the first. Since there are direct projections to and from PFC and preSMA, PFC could monitor preSMA activity for signs of procedural learning.

7.4 Conclusions

In any case in which there are multiple competing systems, there must be some supervisory network that assigns control on each trial to one (or both) of these systems. Although an enormous amount of research has been directed at establishing that humans have multiple learning systems and at identifying qualitative properties of these putative systems, only a few studies have had the goal of studying this supervisory network. Even so, the results presented here suggest some strong conclusions are nevertheless possible.

[8]Despite its name, evidence suggests that the preSMA operates more like an extension of the PFC than like a motor or premotor area (Akkal et al., 2007).

First, the present results strongly suggest that all current models of system switching can be rejected. The main problem is that existing models assume no switch costs and predict better performance than occurs in benchmark tests 2 and 3. Second, the present results also suggest that switching from explicit-rule to procedural responding requires two conditions: 1) poor explicit-rule performance; and 2) good procedural performance. The sparse nature of the existing database makes it difficult to draw any stronger conclusions. Hopefully, the results presented here will inspire researchers to study this important problem in more detail.

Open Practices Statement

The data analyzed in this article were all previously published.

Declarations

Funding. No funding was received for the research described in this article.

Conflicts of interest/Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. Not applicable.

Code availability. Code that fits the GLC and 1D classifier was first developed more than 30 years ago (Ashby, 1992) and today is widely available. The specific version used in this article is available at:
https://github.com/fgashby/Statistical-Decision-Theory-in-Perception-and-Cognition/tree/main.

Author contributions. The author is solely responsible for all aspects of this research.

References

Akkal, D., Dum, R. P., & Strick, P. L. (2007). Supplementary motor area and presupplementary motor area: Targets of basal ganglia and cerebellar output. *Journal of Neuroscience*, *27*(40), 10659–10673.

Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J., & Poldrack, R. A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. *Journal of Neuroscience*, *27*(14), 3743–3752.

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience*, *26*(9), 2424–2433.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ashby, F. G. (2026). *Statistical decision theory in perception and cognition: [signal] detection & general recognition theories*. MIT Press.

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*(3), 442–481.

Ashby, F. G., & Bamber, D. (2022). State trace analysis: What it can and cannot do. *Journal of Mathematical Psychology*, *108*, 102655.

Ashby, F. G., & Casale, M. B. (2003). A model of dopamine modulated cortical activation. *Neural Networks*, *16*(7), 973–984.

Ashby, F. G., & Crossley, M. J. (2010). Interactions between declarative and procedural-learning categorization systems. *Neurobiology of Learning and Memory*, *94*(1), 1–12.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*(5), 204–210.

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

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). Cambridge University Press.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472–482.

Ashby, F. G., Smith, J. D., & Rosedahl, L. A. (2020). Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Memory & Cognition*, *48*(4), 541–552.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science, Second Edition* (pp. 157–188). New York: Elsevier.

Ashby, F. G., Valentin, V. V., & von Meer, S. S. (2015). Differential effects of dopamine-directed treatments on cognition. *Neuropsychiatric Disease and Treatment*, *11*, 1859–1875.

Ashby, F. G., & Vucovich, L. E. (2016). The role of feedback contingency in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(11), 1731–1746.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(3), 363–378.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77–96.

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188.

Blank, H., & Bayer, J. (2022). Functional imaging analyses reveal prototype and exemplar representations in a perceptual single-category task. *Communications Biology*, *5*(1), 896.

Bortz, D. M., Feistritzer, C. M., & Grace, A. A. (2023). Medial prefrontal cortex to medial septum pathway activation improves cognitive flexibility in rats. *International Journal of Neuropsychopharmacology*, *26*(6), 426–437.

Bortz, D. M., & Grace, A. A. (2018). Medial septum differentially regulates dopamine neuron activity in the rat ventral tegmental area and substantia nigra via distinct pathways. *Neuropsychopharmacology*, *43*(10), 2093–2100.

Broschard, M. B., Kim, J., Love, B. C., Wasserman, E. A., & Freeman, J. H. (2019). Selective attention in rat visual category learning. *Learning & Memory*, *26*(3), 84–92.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: Evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*(6), 1598–1613.

Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70*(6), 983–999.

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449.

Crick, F., & Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex*, *8*(2), 97–107.

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1388–1403.

Crossley, M. J., Paul, E. J., Roeder, J. L., & Ashby, F. G. (2016). Declarative strategies persist under increased cognitive load. *Psychonomic Bulletin & Review*, *23*(1), 213–222.

Crossley, M. J., Roeder, J. L., Helie, S., & Ashby, F. G. (2018). Trial-by-trial switching between procedural and declarative categorization systems. *Psychological Research*, *82*, 371–384.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. New York: Oxford University Press.

Ell, S. W., & Ashby, F. G. (2006). The effects of category overlap on information-integration and rule-based category learning. *Perception & Psychophysics*, *68*(6), 1013–1026.

Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*(4), 749–761.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience*, *31*(37), 13157–13167.

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72–89.

Hélie, S., Turner, B. O., Crossley, M. J., Ell, S. W., & Ashby, F. G. (2017). Trial-by-trial identification of categorization strategy using iterative decision-bound modeling. *Behavior Research Methods*, *49*(3), 1146–1162.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*(4), 1013–1031.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S., & Doya, K. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*(10), 464–471.

Kalra, P. B., Batterink, L. J., & Minda, J. P. (2025). Procedural and declarative knowledge simultaneously contribute to category response selection. *Psychological Research*, *89*(5), 146.

Kita, K., Du, Y., Tran, T., & Haith, A. M. (2025). Switching between newly learned motor skills. *Journal of Neuroscience*, *45*(50), 1311–1324.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Maddox, W. T., & Ing, D. A. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100–107.

Minda, J. P., Roark, C. L., Kalra, P., & Cruz, A. (2024). Single and multiple systems in categorization and category learning. *Nature Reviews Psychology*, *3*(8), 536–551.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Ott, T., & Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, *23*(3), 213–234.

Packard, M. G., & McGaugh, J. L. (1992). Double dissociation of fornix and caudate nucleus lesions on acquisition of two water maze tasks: Further evidence for multiple memory systems. *Behavioral Neuroscience*, *106*(3), 439–436.

Packard, M. G., & McGaugh, J. L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, *65*(1), 65–72.

Pasquereau, B., & Turner, R. S. (2017). A selective role for ventromedial subthalamic nucleus in inhibitory control. *Elife*, *6*, e31627.

Paul, E. J., & Ashby, F. G. (2013). A neurocomputational theory of how explicit learning bootstraps early procedural learning. *Frontiers in Computational Neuroscience*, *7*, 177.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience & Biobehavioral Reviews*, *32*(2), 265–278.

Smith, E. E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 249–264.

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., & Grace, R. C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18*(2), 414–421.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J. T., & Ashby, F. G. (2010). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *36*, 54–65.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25*(2), 447–457.

Smith, J. D., Crossley, M. J., Boomer, J., Church, B. A., Beran, M. J., & Ashby, F. G. (2012). Implicit and explicit category learning by capuchin monkeys (Cebus apella). *Journal of Comparative Psychology*, *126*(3), 294–304.

Smith, J. D., & Ell, S. W. (2015). One giant leap for categorizers: One small step for categorization theory. *PloS One*, *10*(9), e0137334.

Smith, J. D., Jamani, S., Boomer, J., & Church, B. A. (2018). One-back reinforcement dissociates implicit-procedural and explicit-declarative category learning. *Memory & Cognition*, *46*, 261–273.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Tulving, E., & Craik, F. I. (2005). *The Oxford handbook of memory*. Oxford University Press.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717–726.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*(1), 168–176.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*(3), 558–584.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(2), 387–398.