Dissociating Categorization Learning System:
Overconfidence and Variations Learning Rates in Accuracy and Confidence Reports

Jordan Richard Schoenherr and Guy Lacroix
psychophysics.lab@gmail.com, guy_lacroix@carleton.ca
Department of Psychology, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

Abstract

In examination of dual-process models of categorizations, research has typically focussed on the manner in which categorization responses change over time. However, one of the basic assumptions of a prominent dual-process account (COVIS) is that an explicit learning system dominates initial stages of training whereas an implicit learning system dominates later stages of training. In three experiments, we consider the utility of using subjective measures of performance (i.e., confidence reports) to continuously sample from a participant's explicit representation of the category structure while also examining changes in these reports over the course of training. The results of an examination of learning rates and the block at which participants reached a performance asymptote support multiple processes and representations accounts of categorization.

Dual-process models of categorization assume that information is processed by and represented in independent cognitive systems. These models have received support from a variety of sources including experimental studies, connectionist simulations, computational models, and neuroimaging studies (For a review, see Ashby & O'Brien, 2005). Several comprehensive dual-process models have been proposed (e.g., Erickson & Kruschke, 1998; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). The present study, however, will focus on a model based on Ashby, Alfonso-Reese, Turken, and Waldron, (1998) competition between verbal and implicit systems (COVIS). This model assumes that learning within the implicit system is dependent on feedback, and can be modelled using a multidimensional variant of single-detection theory (SDT) referred to as general recognition theory (GRT; Ashby & Townsend, 1986) and an explicit system that generates, tests, and modifies hypotheses of low-dimensional category structures.

COVIS presents two assumptions that are central to the present study. First, categorization decisions are made with a criterion, or category boundary (e.g., Ashby & Gott, 1988). In the context of the GRT, stimuli are assumed to have random perceptual effects defined by a joint probability distribution for each of their stimulus dimensions. With the provision of feedback, the category boundary divides separable or integral stimulus dimensions into discrete regions of a categorical space. If a stimulus consists of values along a dimension greater than those specified by the criterion, it is assigned to one category. If the values are less than that specified by the criterion, it is assigned to another category. Using curve fitting, Ashby and colleagues have demonstrated that by the end of training, participants performance is well-described by an optimal classifier model that employs a category boundary.

The second critical feature of COVIS is that the dominance of a given category learning system is dependent on the stage of learning and the structure of the category that has been acquired (Ashby & O'Brien, 2005). During initial stages of learning, categorization is assumed

to be dominated by an explicit hypothesis-testing system. This system categorizes stimuli by identifying and testing one-dimension (1D) rule-based representations using executive resources and working memory. In parallel with the explicit learning process, a feedback-driven procedural-learning process associates a response category with a given region of perceptual space through stimulus-response mapping permitting the retention of a number stimulus dimensions. As the number of implicit memory traces increases over the course of learning, the activation level of the categorization response within the implicit system eventually exceeds that of the explicit system. Thereafter, the implicit system dominates response selection. When no feedback is provided, or when it is delayed by 2500 ms from the offset of a stimulus, participants have difficulty acquiring information-integration categories (e.g., Maddox, Ashby, & Bohil, 2003). Moreover, as the implicit system dominates response selection, response times decrease due to automaticity with longer response times that are found early in training suggesting competition between the implicit and explicit systems.

Subjective measures of performance such as confidence reports were amongst the earliest tools used in the context of experimental psychology and psychophysics to assess difference between awareness and performance (for a review, see Baranski & Petrusic, 1998). Retrospective confidence reports are typically obtained by having an individual assign a subjective probability (e.g., 60%) to the belief that they have provided a correct response. The degree of correspondence between a participant's mean accuracy when assigning a subjective probability to a response is referred to as *subjective calibration* (e.g., Baranski & Petrusic, 1994). Perfect calibration requires that the proportion correct (e.g., 0.6) and mean confidence are equivalent (60%) whereas miscalibration such as over-/underconfidence represents a bias. Studies of perceptual discrimination and general knowledge (e.g., Baranski & Petrusic, 1994) as well as memory (e.g., Koriat, 1993) have observed systematic deviations in the correspondence between task accuracy and subjective probabilities. These deviations can be attributed to differences in the operations supporting primary decision response selection and confidence processing.

An important assumption of COVIS is that the primary decision is based on a multidimensional model of SDT (GRT) and that subjective confidence is determined from a direct-scaling of this evidence (Ashby et al., 1998). With this in mind, it is critical to note that although SDT-based models of confidence processing (e.g., Ferrel & McGooey, 1980), including that proposed by Ashby et al. (1998) are parsimonious, they cannot readily account for several robust findings in the confidence literature. First, the systematic deviations observed in confidence calibration suggest that a direct-scaling of primary decision evidence might not be the sole mechanism used to generate a confidence report (cf. Pleskac & Busemeyer, 2010). Supporting this possibility, studies have demonstrated that the calibration of subjective assessments of performance has been affected by sources of information other than that provided by the target stimulus (Busey, Tunnicliff, Loftus, & Loftus, 2000; Schoenherr, Leth-Steensen, & Petrusic, 2010) and requires additional operations associated with increased decision response time (DRT) when they are reported (e.g., Baranski & Petrusic, 1998; Schoenherr, 2009). Thus, the present study examines whether response accuracy and subjective confidence 1) differ between training blocks and learning rule, as well as whether 2) their learning rate changes depending on task requirements, and 3) the requirement of confidence increases DRTs.

For the purposes of this study we assume that the degree of correspondence between measures of accuracy and confidence can be used to infer the accessibility of representations and the underlying architecture of categorization processes during different stages of learning. First,

when a performance asymptote is used, confidence should reach criterion prior to accuracy given the flexibility of the hypothesis-testing system and should exhibited a more rapid learning rate. Following from this, we additionally assume that participants should exhibit overconfidence when the category structure is readily verbalizable. Third, the requirement of confidence should also increase DRT if it constitutes a secondary process. Moreover, if the hypothesis-testing system and confidence share the same basis, automaticity of responses should occur more rapidly in the rule-based condition relative to the information-integration condition.

Experiment 1a

Method

Stimuli consisted of Gabor patches varying in terms of spatial frequency and orientation. Replicating the method of earlier studies (e.g., Zeithamova & Maddox, 2007), 40 Gabor patches were created for each category for the training phase using the randomization technique by randomly sampling values from two normal distributions. Stimulus values were rescaled into stimulus dimensions with spatial frequency given by $f = .25 + (x_1/50)$ and orientation given by $o = x2(\pi/500)$. Using these values, stimuli were generated with the Psychophysics Toolbox (Brainard, 1997) using MATLAB R2008 (MathWorks, Matick, MA) with an 85% performance asymptote. After a categorization response was provided and a confidence report was obtained, a feedback signal was presented to indicate a participant's accuracy in completing the task. Stimuli were presented to participants using E-Prime experimental software on a Dell Dimension desktop PC.

Procedure

The classification task procedure involved a training phase, consisting of 10 blocks of 40 replications per category, and a transfer phase consisting of 2 blocks of 40 replications per category. Participants were assigned to either the rule-based (RB) or information-integration (II) category structure. In Experiment 1a, participants were provided with both trial-to-trial and block feedback during the training phase. After feedback was provided, participants reported confidence on a 50 (guess) to 100 (certain) scale. In Experiment 1b, only trial-to-trial feedback was provided and in Experiment 2 trial-to-trial feedback was delayed by 2500 ms and the duration of the confidence report.

Figure 1a.

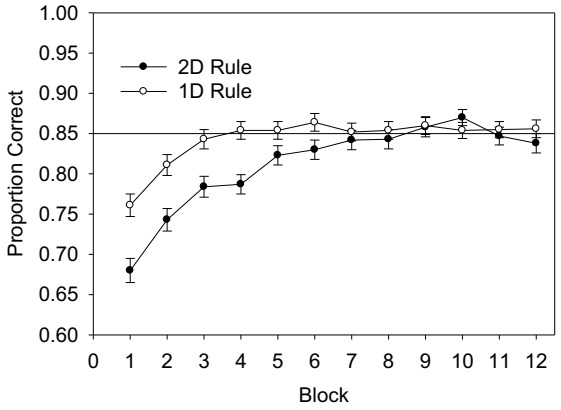

Figure 1b.

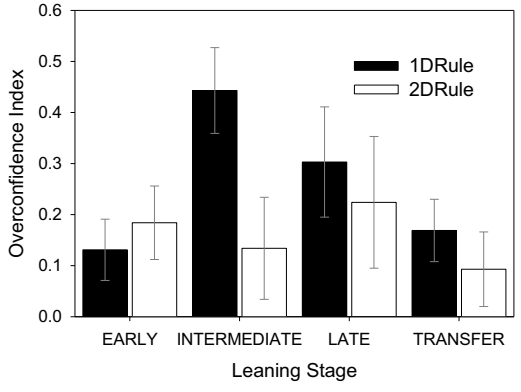

Figure 1. Categorization accuracy (Figure 1a) and overconfidence (Figure 1b) across learning and transfer blocks for 1D (white) and 2D rules (black).

Results

Proportion Correct. As demonstrated in Figure 1a, the results of categorization accuracy replicated earlier findings: 1D rules were learned in fewer blocks then 2D rules, $F(1, 83) = 6.317$, $MSE = .039$, $p = .014$, $\eta^2_p = .071$, and accuracy increased with the number of experimental blocks, $F(11, 913) = 49.167$, $MSE = .005$, $p < .001$, $\eta^2_p = .372$, as well as their interaction, , $F(11, 913) = 6.891$, $MSE = .005$, $p < .001$, $\eta^2_p = .077$. We also found that the requirement of confidence affected category learning as it interacted with block, $F(11, 913) = 2.093$, $MSE = .005$, $p = .052$, $\eta^2_p = .025$. Although the requirement of confidence initially produced reduced performance in the first block ($M = .703$, $SD = .140$) relative to no confidence ($M = .738$, $SD = .131$), participants who reported confidence in the transfer phase were more accurate ($M = .866$, $SD = .112$) then those who did not ($M = .829$, $SD = .103$).

Decision Response Time. Prior to conducted the ANOVA on decision response time (DRT), outliers three standard deviations above the mean were first removed. This accounted for 2.1 % of the total data. As in the accuracy results, participants response time decreased from early to

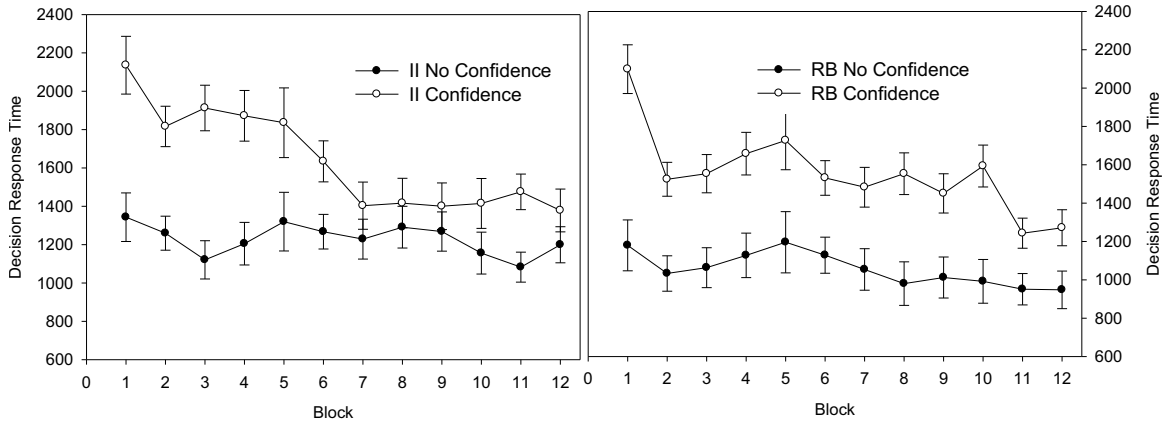

Figure 2. Decision response times with and without the requirement of confidence reports in the rule-based and information-integration learning conditions.

later blocks of training, $F(11, 913) = 11.792$, $MSE = 335411$, $p < .001$, $\eta^2_p = .096$. Of considerable interest to the present study, we found the interaction of experimental block and confidence condition to be significant, $F(11, 913) = 4.600$, $MSE = 335411$, $p = .001$, $\eta^2_p = .039$. As is clear from Figure 3, longer response latencies were observed in conditions when confidence was required. Whereas DRT is relatively constant in the no confidence condition, two rapid rates of change are observed in the DRT when confidence is required between Block 1 and 2 as well as between Block 6 and 7 for the rule-based and information-integration conditions, respectively. This pattern of automaticity suggests that confidence reports required greater additional processing during the categorization response for information-integration category structure than for rule-based category structures.

Confidence Reports. Overall, we found that the overconfidence bias differed across the learning phases (see Figure 1b), $F(1,77) = 8.842$, $MSE = .085$, $p = .004$, $\eta^2_p = .103$. As expected, learning phase was also found to interact with category structure, $F(1,77) = 4.539$, $MSE = .085$, $p = .036$, $\eta^2_p = .056$. As can be seen from Figure 4, overconfidence remained relatively constant in the information-integration condition suggesting that, in general, participants did not have access to the representation that guided their performance. In contrast, an increase in overconfidence was observed in intermediate phases of training in the rule-based condition. This finding suggests that once participants identified the one-dimensional rule, they expected to have continual improvements in performance.

Discussion

The results of our experiment replicate several earlier studies within categorization and confidence processing literature. Categorization performance was found to be affected by the nature of the category structure participants are required to learn: Participants who were required to learn the 1D category structure reached a performance asymptote earlier than those who were required to learn the 2D category structure. Moreover, we also observed that response latencies decreased in fewer blocks for those learning the 1D structure relative to the 2D category structure indicting that participants could more readily acquire a stimulus-response mapping for rule-based categories relative to information-integration categories. Findings such as these conform to the predictions of dual-process accounts of categorization such as COVIS (Ashby et al., 1998) and thereby allow us to interpret the results obtained from confidence reports.

Our analysis of confidence reports also provides evidence for a dual-process account. In the experiment conducted here, we observed increased overconfidence in intermediate phases of training for those participants learning a 1D category structure relative to those who learned the 2D category structure. In general, miscalibration observed here indicates that the representation used to report subjective confidence and that used to respond to exemplars were informed by different sources of information. Greater overconfidence suggests that the category structure that participants were explicitly aware of did not contain the stimulus variability evidenced in the distribution of exemplars. Although it is possible that during the process of rescaling primary decision accumulated evidence could have decayed, it is less clear how this could have resulted in the increases in confidence that lead to the overconfidence observed in our data. More plausibly, it would seem to be the case that different stimulus representations of the category structure were available to two categorization systems.

Table 1. Mean difference between confidence and accuracy for slope and performance asymptote functions across experimental conditions.

Condition	Experiment/Rule	Slope Difference	85% Asymptote Block
Block Feedback	EX1a (1D)	.007	0.23
	EX1a (2D)	-.012	0.45
No Block Feedback	EX1b (1D)	.019	0.12
	EX1b (2D)	.007	0.25
Delayed Trial Feedback	EX2 (1D)	.003	0.36
	EX2 (2D)	.006	0.42

Due to space limitations additional preliminary results of additional experiments could not be reported in great detail. Table 1 contains differences in growth rates for mean accuracy and mean confidence data for the two category structures with the removal of block feedback (Experiment 1b) and when feedback was delayed (Experiment 2). Logistic functions were fit for mean participants performance allowing us to obtain the rate of growth (slope) and a measure of the estimate at which point participants typically reached the performance asymptote (85% correct). As is clear from Table 1, there were generally greater observed positive differences between the slope and performance asymptote for accuracy and confidence functions in the 1D condition relative to the 2D when participants were provided with full feedback (Experiment 1a) or trial-to-trial feedback (Experiment 1b). This suggests that the category structure participants were explicitly aware of was acquired more rapidly than the representation used to inform categorization responses. Supporting this interpretation, differences in the time taken to reach the performance asymptote was much greater in the 2D condition relative to the 1D condition suggesting that participants estimated that they had reached 85% much earlier then they in fact did. Moreover, this difference was again greater when feedback was delayed. Taken together with the analysis of accuracy and confidence indices, we take these findings to suggest two ostensibly distinct category learning and representation systems.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105, 442-481.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, 9, 83-89.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *JEP: LMC*, 14, 33-53.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154-179.

Baranski, J. V., & Petrusic, W. M. (1994). The Calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics*, 55, 412-428.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *JEP: HPP*, 24, 929-945.

Brainard, D. H. (1997). The Psychophysics Toolbox, *Spatial Vision* 10, 433-436.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, 7, 26-48.

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

Ferrel, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behaviour and Human Performance*, *26*, 32-53.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100, 609-639.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *JEP: LMC*, 29, 650-662.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352-369.

Pleskac, T.J. and Busemeyer, J.R. (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, 117, 864–901.

Schoenherr, J. R. (2009). *Mapping Internal Representations of Confidence onto Scales Varying in Range, Interval, and Number of Response Categories*. Unpublished Manuscript.

Schoenherr, J. R., Leth-Steensen, C., & Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. *Attenion, Perception & Psychophysics*, 72, 353-368.

Zeithamova, D., & Maddox, W. T. (2007). The role of visuo-spatial and verbal working memory in perceptual category learning. *Memory & Cognition*, 35, 1380–1398.