Dissociating Implicit and Explicit Category Learning Systems using Confidence Reports

Jordan Richard Schoenherr

A thesis submitted to

the Faculty of Graduate and Postdoctoral Affairs

in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy

in

Psychology

Carleton University

Ottawa, Ontario

Abstract

Dual-process models of categorization (e.g., COVIS) have relied mostly on double-dissociation paradigms and participants' classification accuracy to highlight differences between explicit and implicit modes of learning. In these models, the implicit system uses procedural learning in the absence of attention whereas the explicit system uses hypothesis-testing requiring attentional resources. These accounts assume that the explicit system dominates early stages of learning whereas the implicit system dominates later stages of learning. Thus, differences in response accuracy over the course of learning and between category structures are taken as evidence for explicit and implicit processes. In four experiments, I will consider the utility of using subjective measures of performance (i.e., confidence reports) to continuously sample from participants' explicit representation of the category structure while also examining changes in these reports over the course of training. In Experiment 1, participants were presented with stimuli using the randomization technique using either a rule-based or information-integration category structure and provided with trial-to-trial and block feedback. Block feedback was removed in Experiment 2. In Experiment 3, feedback was delayed to interfere with the implicit learning system while leaving the explicit learning system unaffected. Finally, in Experiment 4, the performance asymptote was lowered to increase overconfidence in participants' performance. Importantly, I observed systematic biases in the relationship between accuracy and confidence reports across training. Confidence reports were more closely associated with explicit representations, produce significant overconfidence for rule-based category structures but only marginally overconfidence for information-integration category structures. These results have important implications for both models of categorization and confidence reports.

Keywords: Categorization, Randomization Technique, Implicit Knowledge, Confidence Processing

Acknowledgements

This thesis represents the culmination of years of intellectual and financial investment on my part as well as many others. I must first acknowledge random genetic variation for giving me an interest to explore the natural and artificial worlds and appreciate both their simplicity and intricacies. More concretely, thanks to my mother's penchant for purchasing encyclopaedias containing the riddles of human culture that have come to preoccupy me, my father for his interest in technology and electronics, and James Jr. whose pulp fiction novels sparked my interest in reading and debate. To my dear sister, Mia, and brother-in-law, Stephen, I owe time and space as well as understanding. My brother Brad has cursed me with my sense of humour, something that has served me well over the years of study. My cousins Mathieu, Jason, and Catherine are to be commended for tolerating my eccentricities as well as attempting to socialize me at times when I had become far too academic. My innumerable nieces (Madison, Serena, and Alisa) and eminently numerable nephew (Erick) have provided the best aid in understanding cognitive development, monitoring and control, as well as what the adult loses.

The academic content of this thesis has as many people to thank. First and foremost my former supervisors Bill and Craig without which I would not have been exposed to the academic culture and ethos that proved a great enough force to incur the opportunity cost that is graduate school. Guy's more recent notable contributions include helping me understanding the practicalities of the scientific culture in which I found myself, indulging my countless digressions, and providing me with the whetstone on which my intellectual instruments have been sharpened. Not so silent partners such as John and covert assistance from Etelle provided a constant source of invaluable insight. Finally, too many lab mates including Rob, Lisa, Tyler, Glen, and Mark have provided tangible and intangible assistance throughout the years.

Table of Contents

Tables and Figures

Appendices

1.0 Introduction

The acquisition, development, and use of categories have been an enduring topic of study across psychology since its inception. They have been the focus of much theoretical and empirical work in disciplines as varied as concept acquisition (Bruner, Goodnow, & Austin, 1956; Hull, 1920; Murphy, 2002), the developmental study of children's ontologies (Carey, 1985; Poulin-Dubois, 1999), speech perception (Liberman, Harris, Hoffman, & Griffith, 1957; Pisoni, Aslin, Perey, & Hennessy, 1982), folkbiology (Medin & Atran, 1999), colour naming systems (Rosch, 1975; Robertson et al., 2000; Kay, 2005), clinical diagnoses (Dawes, 1994; Rosenhan, 1973), gender studies (Herdt, 1994), and artificial intelligence (Anderson & Betz, 2001; Hintzman, 1986). This rich body of work has yielded the progressively clearer understanding that category learning relies on two interactive, yet dissociable information processing systems. The first is a fast-learning, resource-limited explicit system that can learn in the absence of feedback and the second is a slow-learning, high-capacity implicit system that is feedback-dependent (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Krushke, 1998; Nosofsky, Palmeri & McKinley, 1994).

A review of the literature on categorization suggests that a representational shift occurs from early to late stages of learning: early stages of categorization rely mostly on an explicit representational system whereas the later stages rely on an implicit representational system (e.g., Ashby et al., 1998). Traditionally, this research has evaluated the participants' categorical knowledge using measures such as response accuracy by examining the proportion of exemplars correctly categorized (e.g., Ashby, Maddox, & Bohil, 2002; Ashby, Queller, & Berretty, 1999), by calculating the number of trials to criterion (e.g., Maddox, Ashby, Ing, & Pickering, 2004), by estimating the location of a decision boundary in a multidimensional space (e.g., Ashby & Gott,

1988), or by examining the percentage of participants fit by specific decision boundary models (e.g., Ashby & Maddox, 1992). Consequently, these approaches are restricted to drawing conclusions about the system that dominates response selection at any given time while they typically neglect the non-dominant system (although see Regehr & Brooks, 1993; Lacroix, Giguère, & Larochelle, 2005). If humans use two functionally dissociable categorization systems that independently process stimuli to generate categorical responses, then it should be possible to measure both systems' outputs separately.

Subjective ratings, or confidence reports, have been used in a variety of research paradigms to assess awareness of performance (e.g., Nelson & Narens, 1990; Zelazo, Moscovitch, & Thompson, 2007). There is some evidence that these post-decisional reports are based on explicit cues (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Dawes, 1980; Koriat & Ma'ayan, 2005) and could therefore be used to assess subjective awareness of the contents of a nondominant representation. The goal of this thesis will be to investigate whether confidence judgments might assess the representations that are generated by the explicit, hypothesis-testing system throughout the category learning process. If an implicit procedural-learning system does compete with an explicit verbal system to eventually dominate response selection (e.g., Ashby et al., 1998), then changes in the relationship between accuracy and confidence reports should be observed over the course of category learning. Moreover, given that one-dimension (i.e., rule-based) and multidimensional (i.e., information-integration) category structures are thought to be acquired by different categorization learning system (Maddox & Ashby, 2004), I additionally predict that the properties of categorical representations acquired by participants will result in differing levels of subjective awareness due to the relative accessibility of these learning systems.

To investigate these research questions, this present thesis reports four category learning experiments. They all used Ashby and Gott's (1988) randomization technique, a well-established classification task, in conjunction with the requirement that participants provide trial-by-trial confidence reports (Lichtenstein & Fischhoff, 1977). Experiment 1 developed the basic methodology. Participants learned to classify sinusoidal gratings (i.e., "Gabor patches", See Figure 2) using trial- and block-level categorization feedback over multiple blocks of trials. Rule-based and information-integration category structures were both used because they have been shown to solicit the explicit and implicit learning systems, respectively. Following the categorization of the stimuli, participants were then asked to assess their response certainty. Because the category structures were designed to prevent the participants' classification accuracy from exceeding 85%, the participants could be overconfident (i.e., reported confidence is higher than the classification accuracy warrants) or underconfident (i.e., reported confidence is lower than the classification accuracy warrants). Experiment 2 replicated Experiment 1 while removing block-level categorization feedback to establish the reliability of Experiment 1 results. Experiment 3 further clarified the relations among category learning, category structures and confidence by delaying categorization feedback. Delayed feedback been found to affect implicit, but not explicit learning systems. Finally, in Experiment 4, the performance asymptote was lowered to 65% accuracy to examine whether additional negative feedback would affect the explicit representation of the category structure. To anticipate, the results of these four experiments will once more support the assertion that two dissociable learning systems are involved in the categorization of stimuli. More importantly, however, they will establish that confidence reports can be used to probe the nature of the mental representations acquired during the category learning process.

In support of this experimental approach, the literature on categorization and confidence processing will be examined and evaluated. I will begin with a selective review of models of categorization, which will conclude that dual-process models more adequately account for the available behavioural and neurological classification data. Next, I will establish that the literature on confidence judgments and category learning have reached similar conclusions with regard to the nature of the cognitive processes involved in these respective tasks. Confidence judgment theorists have also concluded that the empirical data in this area of inquiry can only be explained by a dual-process account. This review will ultimately lead to the thesis's main hypotheses and the experiments.

2.0 Models of Multiple Category Representations

A considerable amount of research has attempted to describe and explain the cognitive processes that allow people to categorize objects. The processes explored include the use of rules and hypothesis testing (Bruner, Goodnow, & Austin, 1956; Feldman, 2000; Hull, 1920; See also Smith & Medin, 1981), summary representations and prototypes (Hampton, 1979; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973; Rosch & Mervis, 1975; Smith & Minda, 1998), particular instances and exemplars (Estes, 1986; Hintzmann, 1986; Medin & Schafer, 1978; Nosofsky, 1984; 1986; Shin & Nosofsky, 1992), and category boundaries (Ashby, Boynton, & Lee, 1994; Ashby & Gott, 1988; Ashby & Maddox, 1990). All these single process models provide many accurate predictions of response times and categorization accuracy for a wide variety of stimuli (e.g., dot-patterns, geometric figures and feature lists) and category structures (e.g., conjunctive and disjunctive rules).

Despite their successes, single process models have several difficulties. On theoretical grounds, these accounts ignore the influence of prior knowledge on the acquisition of category structure (e.g., Carey, 1985; Gelman & Markman, 1986; Komatsu, 1992; Murphy & Medin, 1985) as well as the distinction between analytic and non-analytic processing (Brooks, 1978; Jacoby & Brooks, 1984). A failure to distinguish modes of processing is problematic given the number of early studies that observed differential contributions of controlled and automatic processes (Schiffrin & Schneider, 1977; Logan, 1988) and the competition and interaction of response systems (see also, Norman & Shallice, 1982). I will return to this below.

In contrast to single-process models, dual-process models assume that categorical information is processed by, and represented in, independent cognitive systems. This fundamental assumption about multiple representation systems is in line with research in multiple major areas of research in cognitive psychology including memory (Hasher & Zacks, 1979; Fernandes & Moscovitch, 2000), reasoning (Evans et al., 1983; Stanovich, 2004; Kahneman, 2011), as well as perceptual and motor decision-making (Fitts & Posner, 1967; Logan, 1988; Norman & Shallice, 1980; Schiffrin & Schneider, 1977). Rather than a task requiring the involvement of either an explicit or implicit process, recent accounts have emphasized the contributions of both systems (e.g., Jacoby, 1991). Support for dual-process accounts of categorization comes from a variety of studies including experimental studies, connectionist simulations, computational models, and neuroimaging studies (e.g., Anderson & Betz, 2001; Ell & Ashby, 2006; Nosofsky, et al. 1994; Smith, Patalano & Jonides, 1998; Vandierendonck, 1995; for an overview, see Ashby & O'Brien, 2005).

Several comprehensive dual-process models have been proposed (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky et al., 1994). They typically assume that an initial

process attempts to generate a simple categorization rule. If feedback disconfirms this initial category structure, another rule is generated. Failure of these simple rules to adequately categorize stimuli leads to the recruitment of a secondary representational system to resolve the category structure. One such model, RULEX (Nosofsky et al., 1994; see also Erickson & Kruschke, 1998), assumes that participants initially consider a single stimulus dimension. If rules based on a single stimulus dimension are not sufficiently diagnostic, the exceptional exemplar is retained in memory as an outlying member of the category. For instance, if a child is presented with a bat, the morphological feature "Wings" might initially be used to categorize the stimulus as a bird. Upon receiving feedback that the creature belongs to the mammal category, the child would then encode a bat as an exception to the mammal categorization rule. Although RULEX can account for categorization performance in complex category structures (e.g., Shepard et al., 1961), it is restricted to the acquisition of mutually exclusive category structures. Although RULEX does postulate that participants store two types of representations (i.e., exemplars and rules; for an alternative multiple representation account, see Smith & Minda, 1998, 2000), it does not specify whether either of these representations is explicitly available to subjective awareness.

2.1 Explicit and Implicit Category Learning Systems

One dual-process model of categorization that has received considerable attention was proposed by Ashby et al. (1998) and is known as COVIS (COmpetition between Verbal and Implicit Systems). COVIS is defined by two independent learning systems: an explicit, rule-based hypothesis-testing system and an implicit, associative procedural-learning system. Categorization within the implicit system is assumed to rely on a multidimensional variant of signal-detection (SDT) referred to as general recognition theory (GRT; Ashby & Townsend,

1986). Stimuli are assumed to take values along one or many physical dimensions and a decision criterion can be used to differentiate stimuli from different categories. After a response is produced, feedback aids in the refinement of the decision criterion in multidimensional space.

COVIS has been shown to be able to explain participants' classification performances for a variety of category structures. Category structures are defined by either separable or integral stimulus dimensions (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1992). In the simplest case, a *rule-based* structure only requires attending to one stimulus dimension whereas *information-integration* structures require attending to stimulus values along two dimensions simultaneously. For instance, Ashby and Gott (1988) provided participants with rectangles that varied in terms of length and width. A rule-based structure would require that participants only attend to one dimension (length, for instance) whereas an information-integration category structure would require that they attend to both length and width. Given feedback, learners are believed to acquire more precise decision criteria or category boundaries in perceptual space. If a stimulus has a value on a given dimension greater than that specified by the criterion, then it is assigned to that category.

COVIS makes several additional assumptions with regards to category learning that follows a logic similar to that of other models of automaticity (Logan, 1988; for a related model of automaticity developed in the context of motor learning, see Fitts & Posner, 1967). During response selection, the two categorization systems compete for dominance based on the level of representational activation. The verbal, hypothesis-testing system dominates early in the course of the acquisition of the category structure. Its dominance is a product of a greater level of activation as a result of its reliance on attentional and working memory resources. If feedback suggests that this rule is inconsistent with the underlying category structure, this system

generates an alternative single dimension rule and provides another response. Concurrently, the implicit, procedural-learning system engages in stimulus-response mapping between a category label and an exemplar. Due to its stochastic nature and its dependence on feedback, the procedural-learning system does not dominate initially. Over the course of learning, however, additional memory traces are encoded and the retrieval of the stimulus-response mappings becomes faster than that of the hypothesis-testing system. As a result, only one system dominates response selection at any given stage of learning. Thus, the dominant system has a higher probability of selecting a stimulus during the response selection stage of decision-making. Revisiting the previous paragraph's example, the hypothesis-testing system could consider a one-dimensional structure based on length when first presented with a rectangle to categorize. Another rule-based representation based on width could also be evaluated. Over time, feedback would generate an association among particular regions of perceptual space and the correct category labels (e.g., stimuli with a small length and a large width could become associated with a given category). This is the categorical knowledge that will ultimately come to dominate performance.

Ashby et al. (1998) additionally acknowledged that these systems remain jointly active under most conditions. Again, it might be the case that participants have the subjective experience of knowing that they are responding correctly most of the time to the stimuli and that they can verbally define the categories. These verbal descriptions, however, might not fully capture the actual features of the categories. For instance, participants might have an explicit, verbalizable representation of a category structure such as "small length" and "large width" that does not take into account certain exception stimuli. Hence, if the hypothesis-testing system dominates response selection using such a rule that fails to be perfectly predictive of category

membership, then it would result in suboptimal performance. In contrast, participants performance is generally well-described by an optimal classification rule suggesting that an alternative representation dominates response selection. Optimal classification performance would instead suggest that responses are guided by an implicit, nonverbalizable representation of spatial frequency units. Thus, taken alone, categorization performance does not suggest in a decisive manner whether an explicit or implicit representation dominate response selection. This points to the more general concern that the demonstration of multiple information processing systems requires the identification of tasks that only affect one system while leaving the alternative system comparatively unaffected (e.g., Dienes & Berry, 1997; Merikle & Reingold, 1991; Reingold & Merikle, 1988). Following from this fact, an alternative means to assess the contributions of the non-dominant system will be considered after examining the explicit-implicit representational dissociation of COVIS in more depth.

The concurrent activation of both categorization systems leads to a second critical feature of COVIS. If explicit and implicit systems have competing representations, then both representations can potentially be sampled. For instance, Ashby et al. (1998) makes claims about the relationship between categorization and confidence processing which might be used to infer process dissociation (e.g., Ziori & Dienes, 2006, 2008; see also Paul, Boomer, Smith, & Ashby, 2011). Thus, confidence reports are a potentially useful tool if they demonstrate a different pattern of responses depending on whether participants have greater subjective awareness of one category structure relative to another. In the following section, I specify the features of the learning systems within COVIS that suggest different representational properties that can be assessed. The methods used to support COVIS are also considered in order to suggest that an alternative method could be used to provide evidence for process dissociation.

2.2 The Randomization Technique, Process Dissociation, and Neurological Evidence.

The potential limitations of experiments that rely solely on categorization responses as a means to determine subjective awareness outlined in the previous section underscore the importance of considering alternative methods to sample the contents of the hypothesis-testing system. The empirical support for category boundaries and COVIS primarily comes from variations on the randomization technique (Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Maddox & Ashby, 1993) that have used concurrent loads and feedback manipulations to dissociate the systems. I will now consider these paradigms.

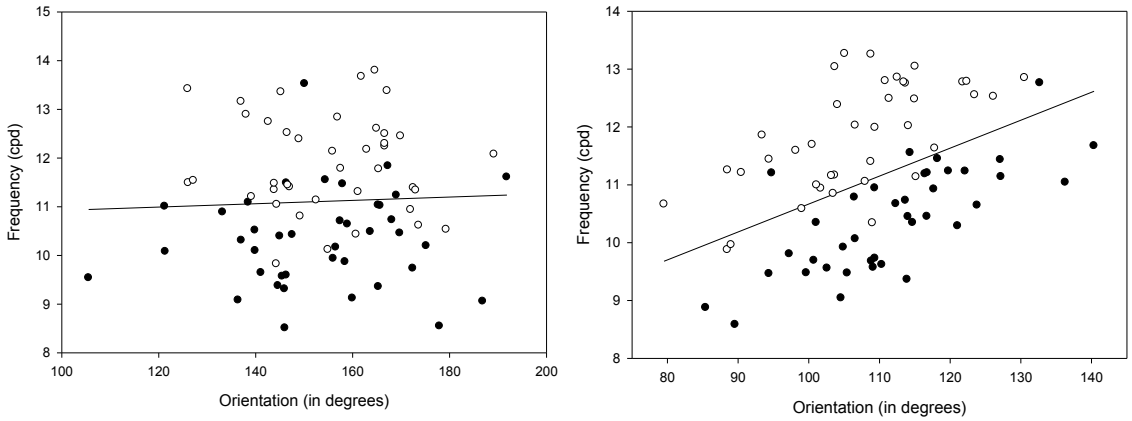

Figure 1. Stimulus distribution for Experiments 4 (High-Overlap Category). Central diagonal represents optimal classification rule.

In a typical randomization technique experiment, exemplars are defined by two normally distributed continuous stimulus dimensions. Categories are defined in this stimulus space in terms of their mean, the variance, and the correlation between stimulus dimensions. For instance, in many experiments, the stimulus dimensions have been the spatial frequency and the

orientation of a Gabor patch (as shown in Figure 2. For other stimulus dimensions, see Ashby & Gott, 1988; Maddox & Ashby, 1993). Under conditions of supervised learning, participants are presented with an exemplar and they must then decide its category membership.

An important feature of the randomization technique is that it allows for category overlap. In other words, it can yield stimulus features that are identical or highly similar for members of opposing categories (i.e., it can yield non-linearly separable categories, see Blair & Homa, 2001, Ell & Ashby, 2006). For example, whales might look like fish and bats might look like birds, but both are in fact mammals. Nevertheless, people do eventually learn to correctly categorize these animals on the basis of other features. Category overlap has proven useful in comparing the validity of different models of categorization.

Methods related to the randomization technique have been used to falsify claims that participants primarily retain exemplar representations of categories. For instance, Blair and Homa (2001) conducted a series of experiments to investigate how people represent linearly and non-linearly separable categories using abstract geometric patterns derived from four category prototypes. In the linearly separable condition, all stimuli within a given category were derived from a single prototype. In the non-linearly separable condition, some category members were selected from the alternative category. Thus, in the nonlinearly separable category, participants had to learn exception exemplars in order to be completely accurate. This did not happen, however. Instead, Blair and Homa found that participants' performance was well described by a linear, optimal classification rule, that did not account for exception exemplars (see also Huttenlocher, et al., 1988; Huttenlocher et al., 1991). Similar evidence has been used to support the claims of COVIS (Ashby et al., 1998). Taken together, this evidence suggests that

participants acquire a decision boundary that divides perceptual space thereby creating two discrete categories.

The results from studies using the randomization technique that provide support for category boundary, and COVIS in particular, have been greeted with theoretical (e.g., Nosofsky & Johansen, 2000; Nosfosky & Krushke, 2002) and empirical challenges (McKinley & Nosofsky, 1995, 1996; Medin & Schwanenflugel, 1981; Smith & Minda, 1998, 2000). For instance, Nosofsky and his colleagues have claimed that changes in attention and stimulus discriminability within a single-process exemplar-based account can adequately model the categorization response associated with dual-process accounts. This suggests that changes in response accuracy, trials-to-criterion, and curve fitting might not be the most effective means to support a dual-process account. Specifically, general models of learning (e.g., Logan, 1988) and categorization (Ashby et al., 1998; Nosofsky, et al., 1994) have suggested that different learning systems will dominate depending on the stage of learning and characteristics of the categories that are being acquired. These claims subsequently motivated the use of additional experimental manipulations.

The suggestion that different category learning systems are postulated to dominate response selection depending on the stage of learning and the type of category structure has been the focus of considerable research using dissociation paradigms (for a review, see Ashby & O'Brien, 2005; Maddox & Ashby, 2004). Evidence for the explicit learning system has primarily come from studies of concurrent working memory loads (Waldron & Ashby, 2001; Zeithamova & Maddox, 2003, 2007; Maddox & Ashby, 2004 for additional sources of evidence; See also Nosofsky & Krushke, 2002). For instance, when participants are required to perform a numerical Stroop Task (Besner & Coltheart, 1979), they have difficulty learning rule-based category

structures, but are relatively unaffected when learning information-integration category structures (Waldron & Ashby, 2001; See also Newell, Dunn & Kalish, 2010). These findings have been taken as evidence representing a *qualitative* change in responding rather than merely a *quantitative* shift in a category boundary location within a single implicit system (Ashby et al., 1998).

In addition to a separable hypothesis-testing system, there is also considerable evidence for an implicit system that is dependent on feedback which has itself been independently modelled (for a review and discussion see Ashby, Ennis, & Spiering, 2007). In an analogous way to classical and operant conditioning, the phylogenetically older procedural-learning system develops stimulus-response associations. In the categorization task, these mappings are many-to-one as multiple exemplars are associated with a category label. Consequently, manipulations that disrupt the acquisition of these stimulus-response mappings should affect information-integration category structure that are learned within the procedural-learning system while leaving rule-based category structures within the hypothesis-testing system unaffected. Ashby, Maddox, and Bohil (2002) provided a clear demonstration of this prediction. They presented participants with one of two training conditions. First, in the observational training condition, participants were asked to pay attention to a category label followed by an exemplar. Ashby et al. predicted that the hypothesis-testing system should be capable of generating a representation of a category structure via observation, but that the procedural-learning system should not because of the absence of feedback. Second, in the feedback training condition, another group of participants was trained with the presentation of exemplars followed by feedback. Under these conditions, the procedural-learning system should benefit more from the provision of feedback whereas the hypothesis-testing system might only exhibit marginal improvements in performance.

In support of their hypotheses, Ashby et al. (2002) observed small performance gains when participants in the rule-based conditions were provided with feedback training relative to observational training. A considerable larger improvement in performance was observed for participants in the information-integration conditions when provided with feedback training, however. Studies have also observed that both positive and negative feedback (referred to as "full" feedback) are required in order to learn information-integration category structures (Ashby & O'Brien, 2007). Moreover, participants have considerable difficulty learning these category structures when no feedback is provided or feedback is delayed by 2500 ms (e.g., Maddox, Ashby, & Bohil, 2003; Maddox et al., 2004; Ashby et al., 1999). In sum, the experiments suggest that, in the absence of other manipulations, the disruption of feedback will affect representation development within the procedural-learning system while leaving those within the hypothesis-testing system relatively unaffected.

Further support for the dissociation of explicit and implicit learning systems also points to the activation of different neuroanatomical areas when participants learn rule-based or information-integration category structures (for reviews, see Ashby et al. 2007; Ashby & Valentin, 2005; Kéri, 2003; Maddox & Ashby, 2004). Evidence suggests that the procedural-learning system is accomplished via activation of regions within the visual associative cortex that project to the striatum which in turn activates units within the prefrontal cortex (e.g., Ashby et al., 2007). In contrast, learning within the hypothesis-testing system proceeds by having the anterior cingulate cortex (ACC) select among alternative category structures corresponding to patterns of activation received by the dorsolateral prefrontal cortex (dlPFC) from the association cortex (for fMRI evidence see, Nomura et al., 2007). In addition to the divergent association of neuroanatomical regions with the respective learning system, the nature of this neurological

model will have important implications for the relationship between categorization performance and confidence reports that likely use explicit representations. As I will discuss in the Conclusion, the similarity in the areas associated with the hypothesis-testing system of COVIS and the executive function that is believed to support error monitoring (Kerns, Cohen, MacDonald, Cho, Stenger, & Carter, 2004; Taylor, Stern, & Gehring, 2007) and presumably confidence processing suggests that these two cognitive systems share many features which differentiate them from the procedural-learning system.

Evidence from studies using process dissociation methods and neurological evidence do much to support the claims of COVIS. These studies, however, are still based on the randomization technique which focuses on the dominant system that produces the categorization response and consequently neglects the contents of the nondominant system. COVIS predicts that participants use information to generate hypotheses concerning simple, unidimensional rules. With repeated exposure to a given stimulus set, participants eventually learn to automatically associate an exemplar with a region in perceptual space (Maddox & Ashby, 1994). Nevertheless, the competitive process between the implicit and explicit systems continues, such that even when the implicit system dominates highly-practiced performance, the explicit system must continue to be activated. Nonetheless, and somewhat surprisingly, the contents of the non-dominant system have not yet been the direct focus of empirical research. For instance, participants' performance is generally assessed in terms of accuracy across learning blocks, the number of participants that can be fit by uni- or multidimensional linear boundaries, or by assessing the number of trials to reach the criterion. Both experimental studies (e.g., Blair & Homa, 2001) and models (e.g., Nosofsky et al., 1994) have suggested that under the same conditions, individual differences might exist in terms of whether participants respond in a

manner reflecting the dominance of either an explicit system or an implicit system when responding. Given that responses obtained using the randomization technique are believed to be principally guided by the dominant system, categorization responses alone do not in themselves present a clear means to assess the representation maintained within the non-dominant system. If the dominant categorization system involved in response selection changes over time, then an alternative measure that consistently samples a given system (e.g., hypothesis-testing system) would help demonstrate a representational dissociation over time.

Hence, the foregoing dissociation studies were based on the premise that a convincing means to demonstrate the presence of the two independent explicit and implicit learning systems is to identify tasks that affect only one of the two systems (Merikle & Reingold, 1991; Reingold & Merikle, 1988). One method that has received increasing interest in categorization studies requires that participants provide subjective confidence reports following their responses (e.g., Balakrishnan & Ratcliff, 1996; Estes, 2004; Rehder, 2003; Rehder & Hoffman, 2005; Ziori & Dienes, 2008). These reports have been used as an alternative means to assess the certainty in category membership (Rehder, 2003; Rehder & Hoffman, 2005; Rehder & Kim, 2009), gradedness of the category structure (Estes, 2004), or more generally, the perceived likelihood that a categorization response was correct (Balakrishnan & Ratcliff, 1996). These studies assume that the information produced by the categorization system can be directly used to inform confidence reports. For instance, Rehder and Kim (2009) obtained results suggesting that mean confidence displayed a pattern corresponding to categorization accuracy. However, using natural (e.g., tree, bird) and artificial categories (e.g., clock, computer), Estes (2004) demonstrated that confidence reports predicted within-category gradedness but not between-category gradedness. Hence, these studies suggest that confidence reports can provide a useful instrument for the

examination of the representation of a category, they leave open the possibility that multiple factors influence confidence over the course of learning or that confidence reports might reflect only features of representations that are accessible in an explicit representation. If COVIS is correct, then it can be expected that participants awareness will change depending on the category structure that is used (i.e., rule-based or information-integration) as well as the stage of learning. Prior to providing a full set of prediction concerning the relationship between categorization and confidence processes, I will consider the processes believed to support confidence reports.

3.0 Confidence Processing and Dissociation of Explicit and Implicit Category Representations

There is good reason to believe that a confidence report methodology is a reliable means to examine subjective awareness. Subjective measures of performance such as confidence reports were amongst the earliest tools used in experimental psychology to assess difference between awareness and performance (e.g., Henmon, 1911; Pierce & Jastrow, 1884; Sumner, 1898). Retrospective confidence reports are typically obtained by having an individual assign a numeric value, or subjective probability, 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* (see Baranski & Petrusic, 1994 and Appendix). Although alternative approaches have used confidence reports to assess explicit awareness in categorization tasks (e.g., Balakrishnan & Ratcliff, 1996; Ziori & Dienes, 2008), there are theoretical and empirical reasons to use calibration indices (see Appendix).

In order to be perfectly calibrated, participants' proportion correct (e.g., $p(cor) = 0.70$) should be equivalent to their subjective probabilities (e.g., $mean(conf) = 0.70$), with perfect

calibration (Cal.) evidenced in no bias (e.g., Cal. = 0.00). In general, participants are typically miscalibrated, yet they rarely exhibit large biases (e.g., Cal. ≤ 0.10). Importantly though, studies of perceptual discrimination and general knowledge (e.g., Baranski & Petrusic, 1994; Gigerenzer, Hoffrage & Kleinbolting, 1991; Keren, 1991; Kvidera & Koustaal, 2008; Lichtenstein, Fischhoff, & Phillips, 1982) as well as memory (e.g., Busey et al., 2000; Hart, 1967; Heathcote, Freeman, Etherington, Tonkin, & Bora, 2009; Koriat, 1997; Koriat, Sheffer & Ma'ayan, 2002; Yonelias, 2002) have observed systematic deviations in the correspondence between task accuracy and subjective probabilities. There is disagreement, however, as to what the degree of correspondence between these two measures represents.

In summarizing the early literature on confidence calibration, Lichtenstein and Fischhoff (1977) observed that participant reports evidenced under- and overconfidence biases in different experimental conditions: overconfidence is observed when reported confidence exceeds the mean proportion correct whereas underconfidence is observed when mean proportion correct exceeds reported confidence. Starting from the assumption that the direction of under/overconfidence bias was a consequence of the properties of the cognitive operations used to perform a task, Lichtenstein and Fischhoff conducted a series of experiments that controlled for task difficulty. The results demonstrated systematic overconfidence for difficult tasks and underconfidence for easy tasks. This pattern, referred to as the *hard-easy effect* (Gigerenzer et al., 1991; see also Griffin & Tversky, 1992), has also been demonstrated in the context of sensory discrimination tasks (e.g., Baranski & Petrusic, 1994) and appears to be robust across a variety of experimental manipulations (Kvidera & Koustaal, 2008). Alternative accounts have also suggested that, rather than task difficulty, the accessibility of information and mental

operations can account for these findings (Dawes, 1980) conforming to suggested requirements for measurement of dissociable representations (e.g., Reingold & Merikle, 1988).

Support for the dissociable contributions to the primary decision process and confidence comes from studies of memory. These studies have found that participants' confidence is reduced relative to their accuracy with increasing experience with the task, referred to as the Underconfidence-With-Practice effect (UWP; Finn & Metcalfe, 2007; Koriat, Sheffer & Ma'ayan, 2002; Koriat, 1997; See also Scheck & Nelson, 2005). UWP suggests that either accuracy increases across successive responses or that subjective awareness motivated by feedback might reduce an extant overconfidence bias. Thus, if confidence reports rely on varying contributions from dissociable implicit and explicit representations available to the participant, differences in overconfidence bias can be used to infer the conditions under which one categorization dominates response selection. For these reasons, a consideration of the relationship between the primary decision process and confidence processing architecture is necessary to understand how competition between these two systems changes over the course of learning. The following sections consider the relationship of confidence processing to the primary decision, the loci of confidence processing, and the sources of information used to compute confidence (see also, Schoenherr, 2008b).

3.1 Confidence Processing Operations

Confidence processing models can be distinguished on the basis of the reliance of the confidence process on the primary decision, the kind of information used to generate confidence reports, and the temporal locus of processing. I will now consider each of these in turn.

Early models of confidence processing in the context of perceptual decision-making assumed that confidence reports were obtained by automatically rescaling the output of the primary decision process (e.g., Balakrishnan & Ratcliff, 1996; Björkman, Juslin, & Winman, 1993; Ferrel & McGooey, 1980; Norman & Wickelgren, 1969; Treisman & Faulkner, 1984). More generally, direct-scaling models do not make a distinction between the process by which confidence reports were obtained and the kind of information used to generate the reports given their reliance on the primary decision. Direct-scaling models assume that once the primary decision has been completed, participants can simply access this information in order to provide an estimate of their performance. Scaling has been hypothesized to be accomplished using either the validity of the cues that were used as a decision heuristic (Gigerenzer et al., 1991) or using the stimulus strength (e.g., Balakrishnan & Ratcliff, 1996; Pleskac & Buseymeyer, 2010; for additional sequential effects, see Treisman & Faulkner, 1984). The latter accounts are of particular importance for the present study as they have been considered in the context of COVIS (Ashby et al., 1998). Strength-based models of confidence are rooted in signal-detection theory. Participants are assumed to use a criterion for primary decision response selection (i.e., if stimulus value > α choose A, otherwise choose B) with additional criteria used to indicate the confidence response (e.g., if stimulus value > δ, choose guess; if stimulus value > β, choose certain). As information for both the primary decision and confidence report are available simultaneously, confidence processing is said to have a *decisional locus*. In this case, primary decision information is simply rescaled into another representation. How this rescaling occurs in the context of these models is generally left underspecified other than assuming that participants compare strength to a confidence criterion.

Ashby et al. (1998) provided a neurological basis for the rescaling of primary decision information into confidence reports in their description of COVIS. They argued that the prefrontal cortex receives projections from the striatum and that the strength of activation within the striatum is the primary determinant of subjective confidence reports. This view of confidence, according to Ashby et al. (1998) is supported by cortical modulation models (e.g., Frith, Friston, Liddle, & Frackowiak, 1991a, 1991b) wherein representations are modulated by top-down processing similar to Norman and Shallice's (1980) Supervisory Attentional System (SAS) (also see Cooper & Shallice, 2000). Specifically, Ashby et al. (1998) assume that both the verbal and implicit systems generate a strength-based representation of confidence by obtaining the distance of an exemplar from the category boundary (see also Balakrishnan & Ratcliff, 1996). They also assume that the procedural-learning system comes to dominate later stages of training. Consequently, in the later stages of learning, the same implicit category representation must inform both categorical responses and confidence reports in the information-integration condition Although Ashby et al. (1998) note that category representations are "assumed to be abstract representations," (p.448), the tacit assumption of the strength-based confidence processing account that they adopt is that confidence scaling occurs automatically (See Pleskac & Buseymeyer, 2010). On this account, if participants are required to report confidence and these reports are dependent on implicit representations, then confidence reports should not correspond to performance when participants learn a rule-based category structure. This follows from COVIS's assumption that the hypothesis-testing system and procedural-learning systems are independent and that confidence reports are only based on the strength of activation within the striatum. Prior to discussing the implication of this statement, I will first consider insights obtained from alternatives to direct-scaling models of confidence processing.

Although direct-scaling models of confidence 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 may not be the sole mechanism used to generate a confidence report. 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 et al., 2000; Schoenherr, Leth-Steensen, & Petrusic, 2010). Second, increases in decision response times have also been observed with the requirement of confidence reports, thereby suggesting that confidence processing requires additional operations. Third, within a confidence scale, confidence response times have often been found to vary between confidence categories (e.g., Baranski & Petrusic, 1998) suggesting that confidence processing constitutes a unique set of operations. In a recent effort to salvage direct-scaling models, Pleskac and Busemeyer (2010) have argued that a stochastic random-walk model might explain these results. They argued that variable confidence response times (termed *interjudgment time* in the context of their model) might be obtained if evidence is assumed to continue accumulating after primary decision response selection is completed. Although this solution appears to be elegant, it does not account for the fact that the primary decision requires additional time when confidence reports are required unless this process is allowed to occur concurrently with the primary decision (e.g., Baranski & Petrusic, 1998; Schoenherr, 2008a). These findings have motivated a second class of models which allow for a distinction between the loci of confidence processing and the primary decision.

3.2 Confidence Processing Loci

Sequential-sampling decision-making models of evidence accumulation, also identified for convenience as rescaling confidence models, assume that separate processing stages are required to complete a task and to assess performance in completing the task (Audley, 1960; Baranski & Petrusic, 1998, 2001; Juslin & Olsson, 1997; Koriat, Lichtenstein & Fischhoff, 1980; Van Zandt, 2000; Vickers & Packer, 1982). For instance, stimulus information activates an accumulator depending on which response they support. Once a threshold amount of information or speeded deadline is reached, primary decision response selection occurs. Then, a confidence report is generated. Many mechanisms have been proposed to account for these processing stages. They include using the inverse of decision response time (Audley, 1960), decision strength (Ferrel & McGooey, 1980), encoding or retrieval cues (Koriat & Ma'ayan, 2005), and the difference in evidence accrued by the accumulators (e.g., Vickers & Packer, 1982).

These models further assume that evidence used in the primary decision is sampled sequentially in terms of either a random-walk (Audley, 1960; Link, 1992) or stored within accumulators that accrue evidence for the response alternatives (Baranski & Petrusic, 1998, 2001; Vickers, 1979; Vickers & Packer, 1982). Some models have additionally suggested that evidence that supports neither of the available response alternatives, so-called nondiagnostic information, could represent the basis for confidence (Baranski, 1991; Baranski & Petrusic, 1998). In all rescaling models, the accumulated information is used to select a confidence category along a response scale. This process requires additional time and processing resources.

Rescaling models of confidence can further be distinguished based on postulates concerning the onset of the confidence. Whereas some models assume that confidence processing occurs post-decisionally (Audley, 1960; Vickers, 1979; Vickers & Packer, 1982), others suggest that it can also occur concurrently with the primary decision (Baranski & Petrusic,

1998, 2001). Evidence supporting rescaling models has principally been obtained from response time analyses. For instance, Baranski and Petrusic (1998) found that when a speeded deadline was imposed on primary decision response selection, participants' confidence response times increased. Alternatively, when accuracy was stressed, confidence response time decreased. According to Baranski and Petrusic (1998), this pattern of results indicates that participants are computing confidence concurrently with the primary decision under the accuracy condition, but are postponing confidence processing until after completion of the primary decision in the speeded condition (see also Baranski & Petrusic, 2001; Petrusic & Baranski, 2003). More generally, such findings can be interpreted as indicating resource sharing between confidence and primary decision response selection. Supporting such a possibility, patterns of performance similar to Baranski and Petrusic (1998) have been observed when a concurrent executive working memory load is presented in conjunction with the primary task: participants postponed confidence reports until after the primary decision (Schoenherr, 2008a). The number of confidence response categories also seems to affect calibration and decision response time (Schoenherr & Petrusic, 2011). Such findings cannot readily be explained within the context of a single-process account without further assumptions (e.g., Pleskac & Busemeyer, 2010).

3.3 Sources of Information used in Confidence Reports

A final distinction can be made among models of subjective awareness more generally. Allowing for the existence of multiple processes introduces the possibility that multiple representations exist, and that they might all be contributing to both the primary decision and the confidence reports. As noted above, whereas overconfidence has often been viewed as a product of difficulty (e.g., Lichtenstein & Fischhoff, 1977), alternative accounts suggest that

overconfidence arises from differential accessibility of information stored within representational systems (Dawes, 1980; Juslin & Olsson, 1997) or cues associated with primary decision information stored in long-term memory (Gigerenzer et al. 1991). Emblematic of these accounts, Dawes (1980) claimed that overconfidence in general knowledge tasks was a result of participants' subjective impression that this type of information was easily accessible whereas underconfidence in perceptual tasks was attributed to the apparent impenetrability of perceptual processes. More recently, Kvidera and Koustaal (2008) conducted a study wherein participants were asked to make paired comparisons of stimuli (e.g., country names such as China and India) that were given different font sizes (i.e., a perceptual dimension) or that had clearly different geographic sizes (i.e., a conceptual dimension). Participants could then be asked to either make a judgment based on the perceptual or conceptual dimension. Kvidera and Koustaal found that overconfidence was exhibited in conceptual tasks relative to perceptual tasks (although, see their Experiment 3). Thus, participants appear to be using different kinds of information to produce confidence judgments in perceptual and conceptual tasks given that the former yield overconfidence whereas the latter yield relatively well-calibrated responses. This pattern of data is consistent with the operation of an implicit system that does not generate consciously accessible representations for perceptual stimuli as well as the operation of an explicit system that does produce accessible representation for conceptual stimuli.

If categorization accuracy is determined by an explicit or implicit categorization system, the difference between performance and confidence reports would yield measures of subjective awareness (Dienes & Berry, 1997; Ziori & Dienes, 2006, 2008). Namely, explicit knowledge should be reflected in proportional increases in accuracy and confidence whereas implicit knowledge should be reflected in a weak correspondence between performance and subjective

reports. This assumption has also been the basis for the use of confidence to dissociate memory systems associated with recognition and familiarity (for a review, see Yonelinas, 2002). In order to use overconfidence bias as an index of subjective awareness, I will now consider the extent to which representations stored in memory are accessible, and whether an explicit or implicit representation is used to report confidence.

In the context of metamemory, early models assumed that reports of subjective certainty were based solely on the target item retrieved from memory. For instance, Hart (1967) assumed that when asked to indicate their certainty in recalling a word, participants assessed the strength of a memory trace (see also Arbuckle & Cuddy, 1969; Cohen, Sandier, & Keglevich, 1991). Later models of metamemory, however, introduced a distinction between the memory trace retrieved from memory and the information used to determine individuals' certainty in their responses (e.g., Benjamin, Bjork, & Schwartz, 1998; Koriat, 1995; for other dissociative frameworks in the context of recognition memory, see Rajaram, Hamilton, & Bolton, 2002; Yonelinas, 2002). Along these lines, Koriat (1997) has suggested that the ease with which information could be retrieved (i.e., availability) was used to determine subjective awareness. Additional evidence suggested that encoding and retrieval cues were used to determine certainty (Koriat & Ma'ayan, 2005). This distinction also conforms to findings of the differential involvement of effortful processing during encoding and retrieval in memory (e.g., Fernandes & Moscovitch, 2000; Hasher & Zacks, 1979) and the "remember/know" distinction in recognition memory (for a review, see Roediger, Rajaram, & Geraci, 2007). In studies of recognition memory, participants identify that they "remember" stimuli when they have a conscious experience of them whereas they identify that they "know" stimuli when they have an experience of familiarity. Yonelinas (2002) suggests that "remember" responses reflect qualitative

information about a memory trace (i.e., conscious experience) whereas "know" responses reflect quantitative measure of memory strength and are associated with high and variable levels of confidence, respectively.

Models of metamemory can help explain how COVIS's learning systems work. Again, COVIS assumes that categorization within the procedural-learning system requires the retrieval of a representation and that response selection occurs when that process is faster than the execution of an explicit classification rule. Given the evidence from the recognition memory literature, I would expect to observe differences in confidence judgments applied to categorization decisions based on the accessibility of these respective representations. Confidence responses in rule-based conditions should be based on highly accessible representations. Moreover, participants' attention should be on the feature (or the very small set of features) to which the classification rule applies. Due to the limits of selective attention, this representation is not likely to include exceptional exemplars when there is category overlap. By comparison, if participants use an implicit representation to classify the stimuli in the information-integration condition, the low-dimensional explicit representation stemming from the rule-based classification system as well as response feedback might be used by participants as cues to generate and report confidence.

By incorporating the representational assumptions of the metamemory literature with the processing assumptions of the indirect scaling confidence model, differences between these two accuracy and confidence can be interpreted to draw theoretically meaningful conclusions about the nature of the different categorization systems. Whereas cue-based accounts in the metamemory literature do not explicitly address the underlying computational process involved in rescaling, these studies do provide evidence for the idea that multiple sources of information

underlie confidence reports. Moreover, the observation that dissociation of memory processes correspond to specific patterns of confidence reports suggests that the under/overconfidence bias can be used to assess the extent to which categorical knowledge is accessible to participants. Consequently, if hypothesis-testing and procedural-learning systems rely on different representations, and if accuracy and confidence reports are differentially influenced by these systems, then specific predictions can be made concerning the magnitude of overconfidence that will be observed both across experimental blocks and between category structures.

4.0 The Interpretation of Overconfidence Bias: Rescaling and Representational Dissociation

In addition to the computational models of rescaling noted above, the representational change that occurs from primary decision evidence into a confidence report must also be considered in order to determine how the under/overconfidence bias can be interpreted. As was made clear in studies of metamemory, confidence reports appears to tap multiple source of information (e.g., accumulated evidence from the primary decision, encoding and retrieval cues, and nondiagnostic information). If confidence processing differentially weights source of information, then modal representations of a presented stimulus within one learning system will need to be abstracted in order for them to be integrated with the accumulated evidence for the other system. Following this integration, confidence responses can be rescaled. Such a process has been considered in the context of lexical development (Karmiloff-Smith, 1992; Mandler, 1988). For instance, Mandler (1988) has suggested that children engage in an additional stage of perceptual analysis that produces sub-lexical perceptual primitives from information obtained from automatic perceptual processing. Similarly, Karmiloff-Smith (1992) has also provided evidence for a progressive shift from implicit to explicit representations occurring across

multiple domains including language acquisition and numeric cognition, processes associated labelling and categorization. General accounts have also described local, modality-based states of "microconsciousness" that are later integrated into a global stimulus representation (Zeki, 2003; for a related account of non-transitive vigilance, see Dehaene & Changeux, 2005, 2011; Dehaene et al., 2006). In terms of strength-based models of accumulated implicit perceptual information assumed by Ashby et al. (1998), some rescaling must occur. Specifically, I would expect that a stimulus-response mapping needs to associate the strength of evidence for the response alternatives with the confidence response key. If multiple representations within the respective categorization systems are allowed, then additional representation integration or weighting stages needs to occur between the categorization stage and confidence response. In either case, a representational change is required. Unfortunately, Ashby et al.'s (1998) strength-based model cannot explain this process.

On such a strength-based account, rescaling primary decision information in order to produce a confidence report would require multiple stages of processing. For instance, such a process could entail processing the sensory signal within an encapsulated nonverbal perceptual learning system. During this process, a spread of activation could occur in long-term memory if the participants have any prior associative knowledge about the stimuli. All available information would then be compared to the accumulated evidence for the response alternatives thereby yielding a magnitude. Once the weight of evidence has been determined for a given comparison, the resulting magnitude would then be mapped onto a confidence response scale. Even if the scaling process can become automatic, the initial mapping of accumulated evidence to confidence categories would require effortful processing. Such a process can account for findings that primary decision response latencies decreases over time (Baranski & Petrusic,

2001) suggesting automatization of the scaling process while also providing a means to explain the observation that confidence response times vary across confidence categories (Baranski & Petrusic, 1998; 2001; Petrusic & Baranski, 2003). As the level of abstraction increases, the sources of information used to assess performance can also be combined allowing confidence reports to be informed by multiple source of information introducing bias (e.g., Schoenherr & Logan, 2013; Ma'ayan & Koriat, 2006). Bias would be introduced in cases where one system dominates response selection (e.g., the procedural-learning system) but another dominates confidence reports (e.g., the hypothesis-testing system). Assuming that confidence reports use information in addition to that used by the dominant categorization system, a comparison between categorization responses and confidence reports can be used to assess the relative contributions of explicit and implicit representations.

The foregoing review leads to several hypotheses. If confidence reports require additional operations and there are systematic biases relating categorization accuracy and confidence reports, miscalibration could be used as a measure of the correspondence between implicit and explicit representations. In general, if a representation for an implicit learning system informs confidence as Ashby et al. (1998) have claimed, then participants should show the greatest calibration and no under/overconfidence bias in their reports when acquiring information-integration category structures. In contrast, participants should show the least calibration and greatest under-/overconfidence bias when acquiring rule-based category structures given that the implicit representation would inform their confidence reports but not their primary decision response selection. The direction of the observed under/overconfidence bias should be suggestive of the relationship between these two category learning systems.

Ashby et al.'s (1998) account also suggests that an implicit representation informs confidence reports. Hence, underconfidence should be observed in rule-based conditions because participants should engage in hypothesis-testing resulting in an explicit, category structure determining response selection. Following from their account, Ashby et al. (1998) need to assume that the activation of the implicit representation determines a participant's response certainty when asked to provide a confidence report. Consequently, this representation should be inaccurate during early stages of training due to the procedural-learning system's feedback dependency. With training, however, the underconfidence bias should be reduced. The implicit categorization system that informs confidence reports would begin to acquire an accurate representation and it would dominate the competition for response selection. In contrast to Ashby et al.'s (1998) strength-based account of confidence, I will instead assume that confidence reports are determined by an explicit representation. On this account, a representation that is accessible within the hypothesis-testing system leads to high response confidence. If the category structure is unidimensional, then participants should rapidly acquire a high degree of calibration in relation to their categorical knowledge. If, however, a performance asymptote reduces the maximum proportion correct that can be obtained I would predict uniform overconfidence. Under these conditions, participants should reach a point in training where their hypothesis-testing categorization system acquires an optimal classification rule. Due to its limited attentional and working memory capacity, the hypothesis-testing system will neither have access to exceptional exemplars nor the proportion of negative feedback available within the procedural learning system. The net effect of an accessible explicit representation of an optimal rule and inaccessible exception exemplar and negative feedback would result in overconfidence. A finding of overconfidence in the rule-based condition would contradict the direct-scaling account

of confidence adopted by Ashby et al. (1998). Given the likelihood that an explicit representation cannot accommodate a large number of exceptional exemplars or stimulus dimensions, overconfidence bias should be an inverse function of the location of the performance asymptote.

If I am correct and the representation generated and maintained within the hypothesis-testing system does indeed inform confidence, I would not expect to obtain underconfidence in the information-integration condition. In opposition to Ashby et al.'s (1998) account of confidence, I predict that participants should initially rely on the accessible explicit representation that provides a unidimensional structure. Under conditions where there is category overlap, participants will likely note that they are continuing to produce a number of incorrect trials. Consequently, confidence should decrease. The combination of increasing accuracy with the dominance of the procedural-learning system later in training as well as decreases in confidence following feedback processing will lead to overconfidence within the information-integration that decreases over time. Crucially, however, the greater influence of negative feedback from incorrect trials that occurs during procedural-learning will mean that the overconfidence expressed within this condition will be less than that observed within the rule-based condition. Increases in the proportion of negative feedback due to decreasing the performance asymptote will increase overconfidence as in the rule-based condition but to a lesser extent given the additional contribution of negative feedback. This general model of explicit representation dominance of confidence reports modified by exceptional exemplars makes specific predictions that will be examined in four experiments that vary category structure, the location of the performance asymptote, and feedback presentation.

4.0 Present Experiments

Previous categorization studies have produced evidence that supports a dual-process account of categorization. These studies have examined the dominant categorization system by using categorization responses but have, consequently, not assessed the contributions of the nondominant categorization system. Dominance has been relative to the category structure as well as the stage of learning. In a series of four experiments, I investigated the effectiveness of confidence reports in assessing subjective awareness of performance to validate this methodology in a categorization task examining multiple category learning systems. In Experiment 1, I established the basic confidence report methodology. In order to replicate the findings of previous studies (e.g., Ashby & Gott, 1988; Ashby et al., 1998) I adopted the randomization technique. The distributions were selected to ensure category overlap (i.e., 15%) leading to a performance asymptote of 85% correct. This created a condition where participants could not obtain 100% accuracy thereby allowing overconfidence to emerge when performance was optimal. Participants were randomly assigned to a rule-based (with one relevant dimension) or information-integration (with two relevant dimensions) category structure. Stimuli appeared one at a time. Participants were required to categorize them as a member of Category A or B and received performance feedback. In addition, some participants were also asked to provide confidence reports following their categorization response (i.e., they provided trial-level confidence). These latter participants received feedback on their classification performance only once they had indicated their confidence. Finally, upon the completion of each training block, participants rated their overall confidence in their performance (i.e., they provided block-level confidence) and they received classification performance feedback for that block of trials.

Accuracy, confidence, calibration indices, and their respective response times for these measures were analyzed.

Experiment 2 examined the possibility that the block-level feedback given in Experiment 1 might have affected participants' confidence ratings on subsequent blocks of trials. Hence, it sought to eliminate this potential concern by removing block-level feedback. The procedure of Experiment 2 was otherwise identical to Experiment 1. This experiment demonstrated the reliability of the paradigm. It also more closely approximated studies of perceptual discrimination where confidence reports have been used previously (e.g., Baranski & Petrusic, 1994).

Given that previous studies have found that delayed feedback reduced performance in information-integration conditions while leaving performance in rule-based conditions unaffected (Maddox et al., 2003), Experiment 3 introduced a delay between categorization response and feedback. I expected that this manipulation would produce performance decrements in the information-integration condition while leaving learning in the rule-based condition relatively unaffected. In the absence of changes to other aspects of the procedure of Experiment 2, a reduction in accuracy was not anticipated to affect confidence reports and was predicted to lead to changes in the amount of observed overconfidence.

Lastly, Experiment 4 examined changes in the performance asymptote used in Experiments 1 to 3. In replication of previous experiments (e.g., Ell & Ashby, 2006), it was known that an increase in category overlap (i.e., 35%) would decrease the performance asymptote (e.g., 65% correct). All other manipulations were identical to Experiment 2. As a result of a reduction in performance relative to Experiments 1 and 2, greater overconfidence was expected to be observed if subjective confidence remained high due to the accessibility of the

representation within the explicit system. I will now consider a set of more specific predictions concerning accuracy, confidence, and their relationship as well as response latencies within Experiments 1 through 4.

4.1 Predictions of Primary Independent Variables.

The design of the four experiments outlined above allows for numerous specific predictions concerning categorization performance and its interaction with confidence reports. In order to establish a baseline, I adopted the randomization technique given its extensive use in categorization of stimuli along continuous stimulus dimensions along with the ability to manipulate category overlap (Ell & Ashby, 2006). Gabor patches were defined along two orthogonal perceptual dimensions (orientation and frequency). After making a response, participants provided confidence reports and received feedback. By using rule-based and information-integration category structures, I assumed that participants would require more trials to learn a one-dimensional, rule-based category structure than a two-dimensional information-integration category structure. Participants in both conditions should reach an identical performance asymptote at which point a direct comparison of confidence reports can be made. The performance asymptote was determined by whether there is a low category overlap (i.e., 15% with an 85% asymptote) or a high degree of overlap (i.e., 35% with a 65% asymptote). In addition to changes in accuracy, I predicted that I would replicate patterns of response time observed in studies of automaticity (e.g., Logan, 1988; Schiffrin & Schneider, 1977). Namely, I predicted that response time would decrease as a function of the number of training trials due to more memory traces associated with stimulus-response mapping. When confidence reports are

required and the representation used to report confidence differs from that used to categorize stimuli, additional processing time should be observed in these conditions.

Predictions can also be made for confidence processing that are dependent of the experimental conditions. In general, mean confidence reports should increase over learning. As participants become accustomed to the paradigm and acquire representations of the category structure, their confidence should increase. In addition, confidence reports should reach an asymptote. If confidence processing draws on the same source of information as the primary decision, mean confidence reports should reflect the 85% performance asymptote. If instead confidence processing is related to the accessibility of the representation, mean confidence reports could asymptote at 100% leading to overconfidence. Participants in the rule-based condition should report higher confidence in their response initially, but those in the information-integration might reach a similar level of confidence later in training. Given the equal proportion of feedback across rule-based and information-integration by the end of training, feedback should only have a significant effect when it is delayed in the information-integration condition due to disruption in acquisition of the category structure within the procedural-learning system.

4.2 Predictions for Miscalibration and Overconfidence.

Specific predictions can also be made concerning the relationship between the implicit and explicit representational systems by examining the correspondence of accuracy and confidence over the course of learning. According to Ashby et al.'s (1998) account of confidence processing, an implicit representation (i.e., signal strength) should inform confidence. If implicit representations provide the basis for confidence reports, participants should exhibit a high degree of calibration in the information-integration condition whereas underconfidence should be exhibited in the rule-based condition. In contrast to this prediction, I will claim that the

differential availability of feedback, the accessibility of the category representations and exceptional exemplars should provide a different basis for confidence reports depending on the category structure and stage of learning. I will therefore consider the specific outcomes within each condition to contrast them with those of Ashby et al. (1998).

If confidence processing uses an explicit representation as well as a secondary set of operations for its generation, I must consider how the combination of sources of perceptual and categorical information could affect the under/overconfidence bias. On a trial-to-trial basis in Experiment 1, participants will receive feedback regarding the accuracy of their responses which will help them build a representation of the category structure. Additionally, it is not likely that participants will maintain representations of exceptional exemplars if they have created an abstract categorization rule when there is category overlap. In the rule-based condition, I assume that participants should have access to the category structure. As noted above, previous findings have demonstrated that the hypothesis-testing system is not as dependent upon trial-to-trial feedback as the procedural-learning system. Similarly, a number of studies using feedback in perceptual discrimination have not observed an effect of feedback on confidence (Björkman, Juslin, & Winman, 1993; Keren, 1988; Schoenherr et al., 2010) or have instead observed inconsistent effects on calibration indices (Baranski & Petrusic, 1994; Petrusic & Baranski, 1997). Consequently, it appears likely that participants should not use feedback to the same extent when using an explicit category representation to report confidence given that the hypothesis-testing system is not dependent on feedback to acquire a category structure.

If feedback does not affect confidence reports that use an explicit representation, I should observe important differences in the under/overconfidence bias between rule-based and information-integration conditions because the category structures overlap. In a typical

supervised learning task, participants learn the properties of the category exemplars through feedback. If there are no exceptions to the category structure, the category representation and feedback convey congruent information – the accessible category representation is a consistently reliable source of information. If exceptions occur as a result of category overlap, however, participants will again create an explicit representation (e.g., an optimal classification rule), but receive feedback that suggests that this representation is somewhat inaccurate. In this case, the complete certainty (100%) with which participants would normally justifiably have in such an explicit representation would be adjusted resulting in reduced confidence. As the hypothesis-testing system does not exhibit as much feedback-dependency as the procedural-learning system, this adjustment in confidence should be less than the total proportion of feedback received (e.g., <15% when there is an 85% asymptote). High certainty in an explicit representation and failure to correctly make adjustments given the proportion of negative feedback would result in overconfidence.

In the information-integration condition in Experiment 1, I would expect the contributions of feedback and explicit representations to be reversed. Namely, given the dependence of the procedural-learning system on feedback, the proportion of negative feedback will likely reduce participants' confidence in their explicit representation to a greater extent than it did within the rule-based condition. Thus, whereas overconfidence should still occur as a result of the accessible explicit representation producing considerable certainty, the negative feedback will reduce overconfidence relative to the rule-based condition. Although some miscalibration and overconfidence is likely to occur given the availability of an explicit representation, the use of response feedback should permit a reasonably well-calibrated assessment of the participants' performance. If, however, participants are greatly influenced by the proportion of negative

feedback, then their level of subjective confidence should be lowered. Moreover, given the UWP effect (e.g., Koriat et al., 2002) I would expect a decrease in overconfidence bias across training. Thus, one of the objectives of the present study is to manipulate the category structure that participants learn in order to alter the category representation as well as change the conditions in which they receive feedback to distinguish between these predictions.

The use of the randomization paradigm in Experiment 1 leaves open the possibility that block-level feedback could alter confidence reports following the first experimental block. Participants could have an explicit representation that determines their confidence, but when told that they have performed better or worse than they had anticipated in a previous block, they might adjust their assessment of performance independently of the representation used during the primary decision process by increasing the magnitude of evidence necessary to obtain certainty. Each block subsequent to the first block of training might therefore be affected by the global feedback provided from the preceding block. Experiment 2 removes block-level feedback, but maintains the conditions of Experiment 1. Consequently, as in Experiment 1, it is predicted that if confidence reports are a function of the hypothesis-testing system, greater overconfidence should be obtained in the rule-based condition relative to information-integration. More specifically, there should be a smaller change in overconfidence bias in the information-integration condition after an initial training phase due to the absence of block-level feedback.

Experiment 3 assesses the effect of trial-to-trial feedback on participants' certainty. Previous studies have demonstrated that the procedural learning system requires feedback in order to form a category representation whereas the hypothesis-testing system does not. Thus, delaying feedback should reduce accuracy in the information-integration condition whereas it should not affect performance in the rule-based condition to the same extent. This has important

implications given Ashby et al.'s (1999) assumptions about confidence. If, as they have suggested, confidence reports are derived from an implicit representation of stimulus strength, participants should still exhibit a high degree of calibration in the information-integration condition despite reductions in primary decision accuracy. In contrast, while participants in the rule-based condition should have greater accuracy relative to the information-integration condition, they should exhibit less overconfidence bias due to reduced confidence as a result of reduced learning in the procedural-learning system that generates confidence.

The alternative account of confidence process as primarily determined by explicit representations and feedback produces a contrasting set of predictions. On this account, the hypothesis-testing system will still acquire an explicit representation of the category structure that can be used to inform confidence reports in both rule-based and procedural-learning conditions. I would again assume that participants in the rule-based condition rely on the explicit system for both categorization judgments and subjective confidence. Even if categorization responses of participants in the information-integration condition use the procedural-learning system to categorize stimuli, their confidence reports should also be determined by the explicit representation within the hypothesis-testing system. The greater quantity of negative feedback that participants in the information-integration condition receive early in training should result in lower overconfidence. If the accessibility of the explicit representation were the only determinant of subjective confidence, I would expect greater levels of overconfidence in the information-integration condition in Experiment 2 then in the previous experiments. However, I will instead assume that the contributions of negative feedback to confidence reports will maintain the same level of overconfidence bias in both rule-based and information-integration conditions during training as those observed in Experiments 1 and 2. This would result from participants in the

information-integration condition using negative feedback to adjust their confidence reports. With the removal of trial-to-trial feedback during transfer, I would additionally predict an increase in overconfidence bias given that in the absence of feedback the only information available is the explicit representations within the hypothesis-testing system.

The results of Experiments 1 through 3 leave open the possibility that participants' level of confidence might not be a result of representational accessibility. Instead, participants might simply have a general bias towards a mean level of confidence such that they assume that they are quite accurate (e.g., 90%). In this case, negative feedback might only be used to adjust predicted performance. If instead confidence reports are affected by the accessibility of the explicit representations, then reducing the performance asymptote should not alter the extent participants' certainty: regardless of a performance asymptote, an optimal classification rule maintained within the hypothesis-testing system would have the same accessibility. Consequently, the degree of calibration evidenced in previous experiments might not be the result of a correspondence between accuracy and confidence reports but could instead be an artefact of the correspondence between the performance asymptote and an internal bias. Alternatively, if confidence reports are based on the accessibility of the explicit representation and feedback, participants should adjust their confidence reports accordingly. If the same degree of calibration is observed in Experiment 4 as in Experiments 1 through 3, this would simply suggest that the proportion of negative feedback is the primary determinant of confidence reports. Under these conditions, participants would simply use the proportion of negative feedback to determine their subjective confidence. If instead greater overconfidence is observed, this would suggest that the explicit representation is the main determinant of confidence reports but that feedback is used imperfectly to adjust the estimate that is initially based on the explicit

representation. Given that the same relationship between explicit and implicit representation should hold in Experiment 4, I would again predict greater overconfidence in the rule-based condition relative to the information-integration condition. I will now describe a general method to allow for these experimental manipulations.

4.3 General Method

The four experiments were conducted using the randomization technique and confidence reports to validate a confidence report methodology as a means to dissociate the representations contained within the respective learning system. Given the similarity of methodology used in each experiment, a general description of stimuli, materials and procedures is provided below. Specific information is provided for each experiment in the appropriate sections.

4.3.2 Stimuli and Apparatus

The 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 (M_x, M_y). The categories were defined by a rule-based (spatial frequency or orientation of line pattern) or information-integration category structure. Stimulus values were transformed into values along the stimulus dimensions in the same manner as Zeithamova and Maddox (2006) 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; Pelli, 1997) using MATLAB R2008 (MathWorks, Matick, MA). Sample stimuli are presented in Figure 2 and the distribution used in Experiments 1-3 is presented in Figure 3. The parameters for this distribution are displayed in Table 1. The transfer phase additionally contained the category prototypes (see Table 1) and

central tendencies of each category. Stimuli were presented to participants using E-Prime experimental software (Schneider, Eschman, & Zuccolotto 2002a; 2002b) on a Dell Dimension desktop PC.

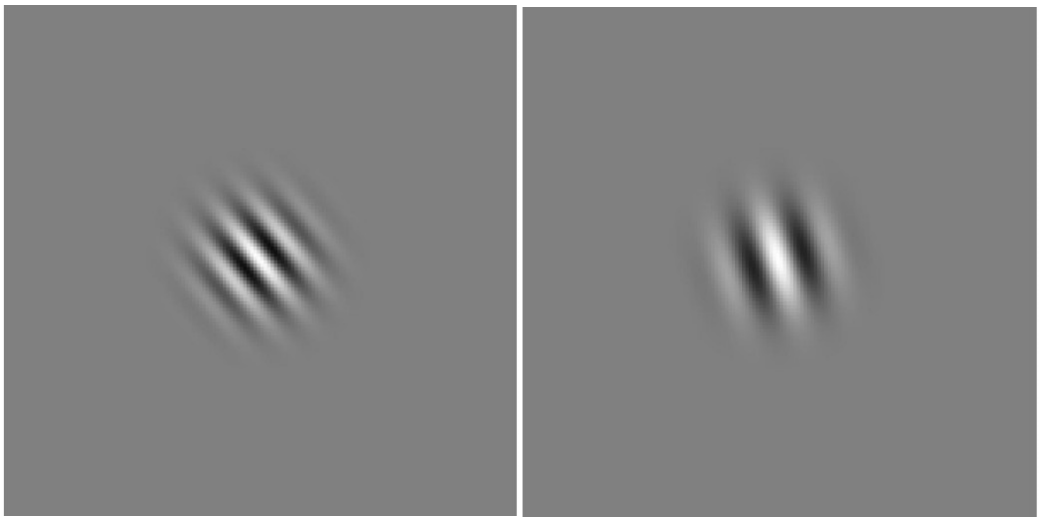

Figure 2. Sample stimuli from Category A and Category B distributions, respectively. Category labels were randomly assigned in order to counterbalance stimulus distributions.

Table 1. Parameter Values Used to Generate the Stimuli in Experiments 1 – 4 at Each Level of Category Overlap. Orientation units are presented in terms of degrees. Spatial frequency units are presented in equivalent units but were converted into cycles per degree (cpd).

Condition	Means				Variance		COVAR
	Spatial Frequency		Orientation		Spatial Frequency	Orientation	
	A	B	A	B			
Experiments 1-3 (Medium)	101.5	118.5	118.5	101.5	162.5	162.5	112.5
Experiment 4 (Medium-high)	106.1	113.9	113.9	106.1	162.5	162.5	112.5

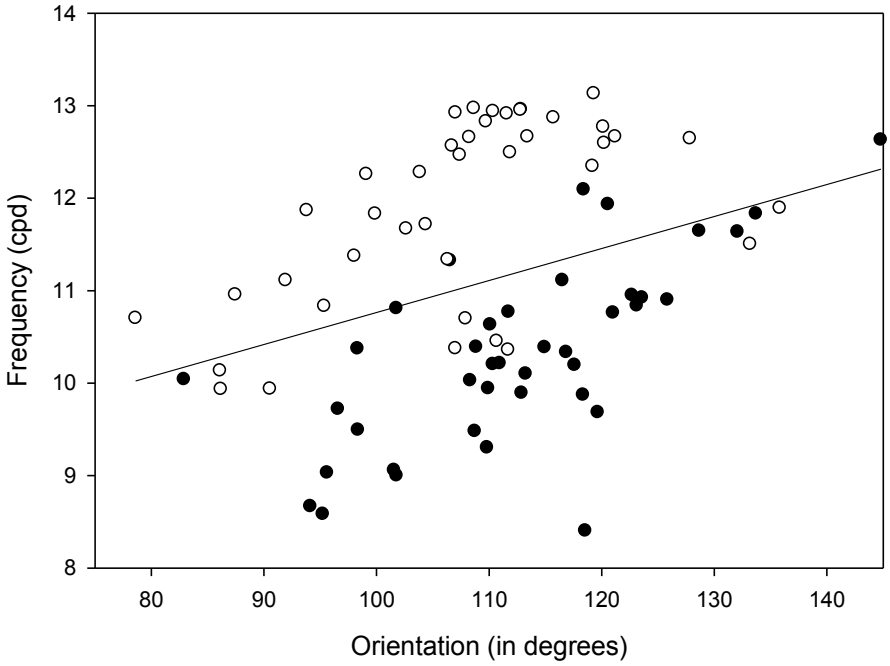

Figure 3. Stimulus distribution for the information-integration condition in Experiments 1-3. The central diagonal represents the categorical boundary.

4.3.3 Procedure

The participants were first given a short memory span test consisting of 10 trials of 8-item stimuli prior to the experiment. Participants were presented with a string of 8 letters for 1,000 ms and then were asked to recall those items in forward or backward order. Recall order was counterbalanced. The results of the memory span test were not analysed in relation with the classification and confidence data for the present thesis. Hence, it will not be discussed further.

All experiments were divided into a training phase and a transfer phase. On every trial, participants were presented with a Gabor patch randomly sampled from the 40 stimuli from either Category A or Category B distributions. Participants were required to press "C" for stimuli

in Category A and “M” for stimuli in Category B. These response keys were counterbalanced. After a response was provided, the stimulus was removed from the screen and the participants were then prompted to indicate confidence in their response. Ratings were provided on a 6-point, 10-unit interval Likert scale with “50” representing a guess and “100” representing certainty. Confidence categories 50 through 100 were assigned to keys “E” through “I”, respectively. The keys were labelled with stickers with the appropriate confidence categories identified on them. These confidence categories were selected on the basis of Schoenherr and Petrusic's (2011) recommendations following from a comparison of various scale parameters. In that study, they observed that scaling biases were the smallest for 6-point scales and, given that such scales have been used in many studies, it appears principled to adopt it here.

After participants reported their confidence, they were provided with feedback about their classification accuracy. Feedback consisted of presenting the stimulus again with corrective feedback beneath ("CORRECT" or "INCORRECT") and an auditory signal. A correct response was indicated with a 500 Hz tone whereas an incorrect response was indicated with a 200 Hz tone. Feedback was presented in both modalities immediately following confidence reports for 500 msec. This manipulation was performed as both positive and negative feedback are required to effectively learn information-integration category structures (Ashby & O’Brien, 2007). An Inter-trial Interval (ITI) consisted of a 2,500 ms pause to allow feedback processing.

In Experiment 1, the training phase consisted of 10 blocks that included 80 trials each. Upon completion of a block of trials, participants were provided with feedback that displayed the percentage of correct categorization responses for the previous block of trials (e.g., 85% correct). They were then asked to provide their confidence that they had learned how to correctly classify instances from Categories A and B. Confidence report used the same 6-point scaled that was

used on a trial-by-trial basis. In Experiments 2-4, participants were simply required to report their overall confidence for the block but did not receive block-level feedback. The first 6 blocks of the training phase were administered consecutively whereas the remaining 4 blocks of training were administered in a run that followed a brief break. In conjunction with a performance asymptote, this quantity of training allowed participants to reach asymptote thereby yielding optimal performance in the transfer phase (Hélie, Waldschmidt, & Ashby, 2010). Finally, the transfer phase consisted of 2 blocks of trials, each consisting of 80 trials from the training phase and 4 additional exemplars (2 category prototypes and 2 category central tendencies). Neither trial- nor block-level feedback was provided during the transfer phase, but in all other respects, the procedure and stimuli used in these blocks were the same as those of the training phase.

Upon completion of the transfer phase, participants were asked to provide a written description of the category prototype. They received the following instruction: "In one sentence, try to describe the line pattern that *best* reflects the typical characteristics of each category. If you are not certain, please guess." Participants also assigned a confidence rating to this description. Finally, the participants were provided with a computer program that allowed them to recreate the visual patterns that they identified as the line pattern that best reflects the typical characteristics of Category A and Category B. Using a script written in Python, participants were presented with a random Gabor patch on the computer selected from the distributions used in Experiments 1-3. Participants used the "up" and "down" arrow keys to increase or decrease the frequency of lines on a Gabor patch and the "left" and "right" arrow keys to change the orientation of the Gabor patches. Once complete, the participants pressed the "Enter" key to save the stimulus parameters and reported confidence. Both outputs were saved to an Excel file. These

data about the participants' knowledge of the category prototypes were not analysed for the thesis and hence will not be discussed further.

5.0 Experiment 1

The purpose of Experiment 1 was to examine the effect of confidence reports on participants' category learning. More specifically, I sought to assess whether these reports could be used as an alternative to double-dissociation paradigms to provide support for dual-process accounts of categorization. Previous research has used confidence judgments to examine response certainty (Paul et al., 2011). This method has also been used to examine the relationships between confidence ratings and classification accuracy (e.g., Rehder, 2003; Ziori & Dienes, 2006). No categorization research, however, has yet used trial-by-trial confidence reports to report calibration indices (e.g., Lichtenstein & Fischhoff, 1977; Baranski & Petrusic, 1994).

Compared to previous studies, the use of calibration indices has several advantages relative to alternative methods for assessing uncertainty. For instance, Paul et al.'s (2011) method requires that participants provide a response that indicates their uncertainty in the category of the stimuli rather than attempting to classify the Gabor patch and then reporting confidence. Under these conditions, participants would not receive feedback given that they did not categorize the stimulus. As a result, their method alters the nature of the category structure participants learn because no feedback is given on uncertain trials as participants have not selected a category. In contrast, the addition of confidence reports following stimulus categorization does not affect the proportion of feedback received while additionally providing information concerning participants' certainty on a trial-to-trial basis. The use of a numeric confidence scale also has the benefit of allowing a direct comparison between obtained accuracy and subjective confidence

(cf. Balakrishnan & Ratcliff, 1996). Unlike Ziori and Dienes (2006, 2008), however, I will consider the full range of confidence reports rather than using a subset of responses from the confidence scale.

A potential drawback of the requirement of confidence reports is the possibility that it will have an effect on category learning. Participants' performance could be improved if confidence reports induce performance monitoring (e.g., Schoenherr & Logan, 2014; cf. Nelson & Narens, 1990). This could happen for different reasons. For instance, performance monitoring could increase accuracy because it might lead participants to allocate their attention more efficiently to the relevant stimulus dimensions associated with competing rules on a trial-by-trial basis. Alternatively, participants might use feedback more effectively to adjust the criteria used for response selection. The requirement of confidence reports would therefore be accompanied by increased accuracy. In contrast, confidence could reduce performance. Even though performance monitoring might occur, confidence processing and the hypothesis-testing system could use the same processing resources. For example, I would expect to observe this result if confidence processing represents a concurrent secondary task that utilizes the same stimulus information and cognitive resources as the primary decision.

By comparing conditions in a group of participants that provide confidence reports to those who provide no confidence reports, I can determine the extent to which their responses are affected by the requirement to report confidence. Block-level confidence reports were also requested to examine whether participants would be better calibrated when assessing their performance over a number of trials. These block-level assessments can be compared between the groups that did and did not report trial-level confidence.

The hypotheses for this experiment are as follows. First, participants' classification accuracy will improve across training blocks, reach asymptote, and maintain itself during transfer. Second, categorization decision response times (DRTs) should also decrease across learning due to the developing automaticity of stimulus-response mappings. Third, in replication of previous research (for a review, see Maddox & Ashby, 2005), participants should learn rule-based category structures more rapidly than information-integration category structures. Fourth, if confidence reports are a product of information retained within the hypothesis-testing system, then confidence reports should exceed response accuracy in both rule-based and information-integration conditions. Nevertheless, the overconfidence bias should be less pronounced when participants are required to learn an information-integration category structure as compared to a rule-based category structure. This prediction stems from the assumption that the former is learned by the procedural-learning system which relies on feedback and an inaccessible implicit representation of the category structure. These predictions can be contrasted with those inferred from Ashby et al. (1998) wherein confidence reports are automatically scaled from information within the procedural-learning system. Moreover, if confidence reports represent an additional process that requires rescaling, the requirement of confidence should increase categorization response time. Given the similarity of the representations used in the hypothesis-testing system and confidence reports, response times in the rule-based condition should be shorter than those in the information-integration condition.

5.1 Method

5.1.1 Participants and Design.

One-hundred and thirty-two participants were recruited from Carleton University and were awarded course credit for their participants in a 2-hour session. Participants were randomly assigned to one of four experimental conditions. They were required to learn a rule-based category structure (defined by frequency; $n = 63$) or an information integration category structure (defined by frequency and line orientation; $n = 69$), and they either performed the standard categorization task providing block-level confidence only ("block-level confidence condition"; $n= 47$) or were required to provide both trial- and block-level confidence reports ("trial-and-block confidence condition"; $n = 85$). Uneven sample sizes were selected due to the need to eliminate participants who only reported block-level confidence when analyzing mean trial-level confidence and confidence calibration indices. As a result, more participants were included in the trial-and-block confidence condition than in the block-level confidence condition. No participants reported hearing or visual impairment.

5.1.2 Stimuli and Procedure.

The stimuli and the procedure were identical to those described in the general methods section. The one exception is that following the completion of a block of trials and reporting of block confidence, participants received block-level performance feedback in the form of a percentage of correct response that they provided within that block of trials (e.g., 62%).

4.2 Results

The results of Experiment 1 were analyzed across training and transfer blocks to examine how accuracy and confidence change over time. In order to avoid the inclusion of participants who failed to learn the categories, those who obtained less than 65% correct over all experimental blocks were excluded from the analyses. The total number of excluded case by their experimental condition is included in Table 1. Following the removal of these participants, I next sought to determine whether participants reached the 85% performance asymptote. A series of one-sample *t*-tests were performed using an 85% correct learning criterion to make this determination. These test revealed that there was no significant differences for the criterion in comparison to the final learning block (M = .85, SD = .10), $t(138) = -.60$, $p = .55$, or the mean for the last two learning blocks (M = .85, SD = .09), $t(138) = -.65$, $p = .51$. Moreover, comparisons of guessing responses (50%) to performance across all blocks as well as learning blocks and transfer blocks revealed that participants had learned to differentiate features of the responses categories well above chance, all ts > 36, all ps < .001.

A series of mixed-design analyses of variance (ANOVAs) were conducted on the following dependent variables: proportion correct, categorization decision response time (DRT), confidence response time (CRT), confidence calibration, and over/underconfidence bias. Where appropriate, Greenhouse-Geisser adjusted values were used. To facilitate comprehension, however, only unadjusted degrees of freedom are given.

Table 2. Excluded participants by experiment and confidence condition based on a 65% guessing criterion (Experiment 4* adopted a 50% guessing criterion due to the reduction in performance asymptote). With the exception of Experiment 3, all experiments included a confidence and no confidence condition.

	Rule-Based		Information-Integration	
	Confidence	No Confidence	Confidence	No Confidence
Experiment 1	0	0	5	1
Experiment 2	0	3	2	2
Experiment 3	0	NA	4	NA
Experiment 4*	0	1	2	3

4.2.1 Proportion Correct

A 2 (Categorization Rule: rule-based vs. information-integration) x 2 (Confidence Condition: block-level confidence vs. trial-and-block confidence) x 12 (Experimental Blocks: 1-12) mixed-design ANOVA was performed on the proportion of correct responses collapsing across response key. The results are shown in Figure 4. As expected, an increase in accuracy was observed across training and transfer blocks, $F(11, 1342) = 50.12$, $MSE = .01$, $p < .001$, $\eta^2_p = .29$. Supporting earlier categorization studies that have used the randomization task (e.g., Ashby et al., 1998), I observed a main effect of Categorization Rule, $F(11, 122) = 11.01$, $MSE = .04$, $p = .001$, $\eta^2_p = .08$. Participants in the rule-based condition were more accurate than participants in the information-integration condition (see Table 2). A significant interaction was also observed between Experimental Block and Categorization Rule, $F(11, 1342) = 8.01$, $MSE = .01$, $p < .001$,

η^2_p = .08. A decomposition of this effect revealed a significant simple effect for the rule-based condition, $F(11,1430) = 11.51$, $MSE = .01$, $p < .001$, as well as for the information-integration condition, $F(11,1430) = 37.79$, $MSE = .01$, $p < .001$. As can be seen in Figure 4, the participants from both conditions were consistently performing at asymptotic levels during the second half of the experiment. The main effect of Confidence Condition was not significant, $F(1, 122) = .44$, $p = .507$.

Table 3.Averaged accuracy, mean confidence, and calibration indices for rule-based (RB) and information integration (II) category structures in Experiments 1-4. Standard errors are presented in parentheses.

Ex.		p(correct)	M_{Conf}	Calibration	O/U
1	RB	.85 (.01)	90.94 (1.21)	.03 (.00)	.06 (.00)
	II	.81 (.01)	87.96 (1.24)	.02 (.00)	.06 (.00)
2	RB	.83 (.01)	92.26 (1.28)	.03 (.00)	.08 (.01)
	II	.82 (.01)	89.76 (1.34)	.02 (.00)	.07 (.01)
3	RB	.84 (.01)	93.45 (1.39)	.02 (.00)	.10 (.01)
	II	.78 (.01)	88.48 (1.37)	.04 (.00)	.10 (.01)
4	RB	.64 (.01)	81.96 (2.36)	.07 (.01)	.16 (.02)
	II	.67 (.01)	78.27 (2.17)	.05 (.01)	.09 (.02)

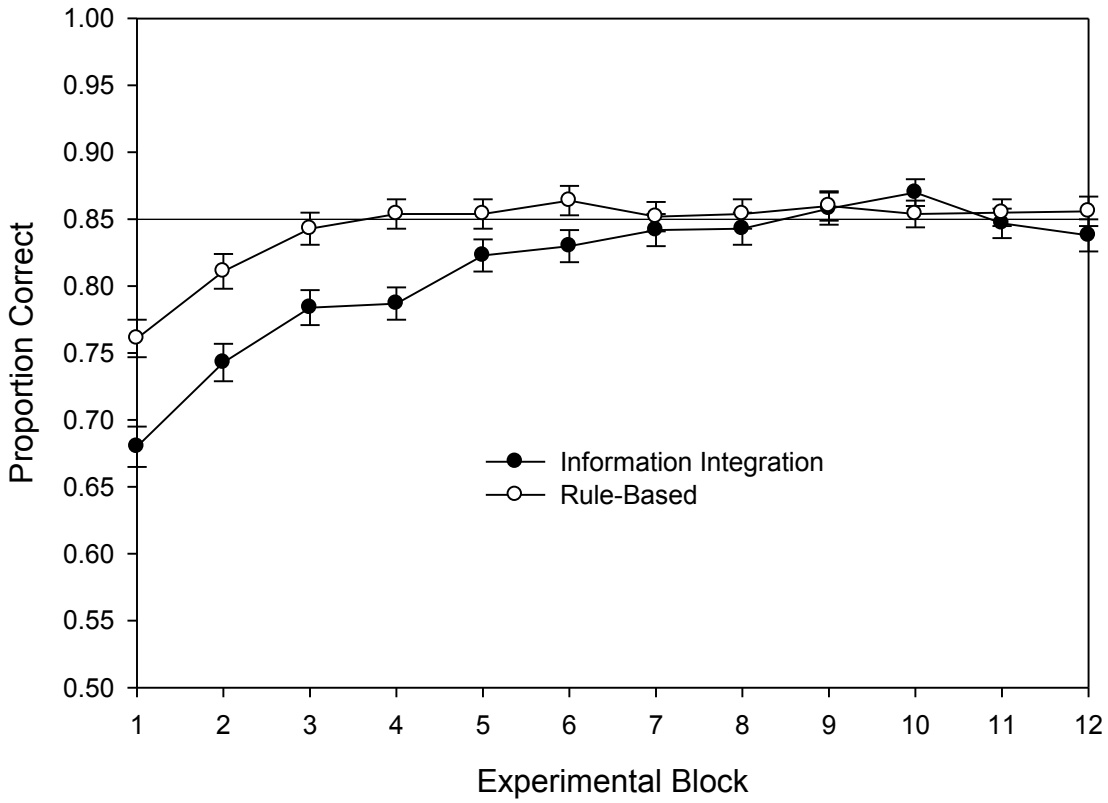

Figure 4. Average accuracy across category rule conditions. The reference line shows the performance asymptote. The error bars represent the standard error of the mean.

4.2.3 Decision Response Time

The design of the categorization decision response time (DRT) analyses was identical to that of the accuracy analysis. Prior to conducting the ANOVA on DRT, outliers three standard deviations above the mean were first removed. They accounted for 2.39% of the total cases. The results are shown in Figure 5. Supporting the patterns of learning evidenced in the analysis of accuracy, participants' DRTs decreased from early to later blocks of training, $F(11, 1364) = 11.21$, $MSE = 339845$, $p < .001$, $\eta^2_p = .08$. This generally suggests a process of automatization

resulting from the consistent mapping of a set of stimuli to a response key (e.g., Logan, 1988; Shiffrin & Schneider, 1977). There was also a marginally significant trend for the responses in the rule-based condition (M = 1291 ms, SE = 50) to be faster than responses obtained in the information-integration condition (M = 1428 ms, SE = 49), $F(1, 124) = 3.79$, $MSE = 1735092$, $p = .054$, $\eta^2_p = .03$. Responses in the trial-and-block confidence condition (M = 1566 ms, SE = 42) were also found to slower than those in the block-level confidence condition (M = 1153 ms, SE = 56), $F(1, 124) = 34.79$, $MSE = 1735092$, $p < .001$, $\eta^2_p = .22$.

Of considerable interest to the present study, a significant interaction was observed between Experimental Block and Confidence Condition, $F(11,1364) = 5.76$, $MSE = 339845$, $p < .001$, $\eta^2_p = .01$. A decomposition of this effect revealed a significant simple effect for the trial-and-block confidence condition, $F(11,1364) = 24.22$, $MSE = 185766$, $p < .001$, but not the block-level confidence condition, $F(11,1364) = 1.35$, $MSE = 185766$, $p = .19$. This pattern suggests the participants' responses latencies differed when confidence was required.

As is clear from Figure 5, longer response latencies were observed in conditions for which trial-and-block confidence reports were required. Whereas DRT was relatively constant in the no confidence condition, a drop in DRT is evidenced in the confidence condition between

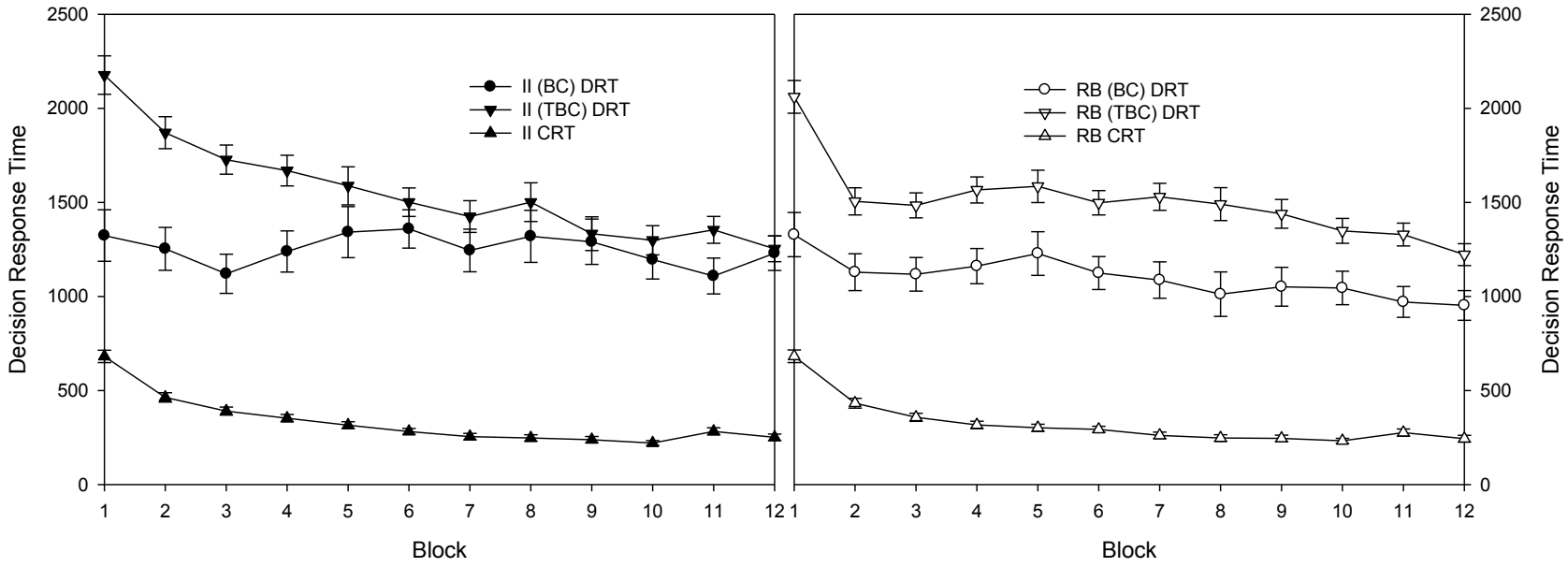

Figure 5. Categorization decision response time (DRT) with trial-and-block confidence (TBC) and with block-only confidence (BC), and confidence response time (CRT) over experimental blocks for rule-based (RB) and information-integration (II) category structures. The error bars represent the standard error of the mean.

blocks 1 and 2 with a DRT monotonically decreasing thereafter. This pattern of automaticity suggests that confidence reports required additional processing and that this occurred during categorization process. A delay during categorization can be taken as evidence for a concurrent confidence processing (e.g., Petrusic & Baranski, 1998). An analysis of CRT presented below will demonstrate that some of this processing occurred post-decisionally following a similar pattern.

4.2.3 Confidence Calibration Indices

4.2.3.1 Mean Trial-Level Confidence. Prior to examining the calibration indices, mean confidence reports were obtained to establish what factors affected confidence. A 2 (Categorization Rule: rule-based vs. information-integration) x 12 (Experimental Blocks: 1-12) mixed-design ANOVA was used in the analysis of categorization trial-level confidence reports. As a portion of the participants was assigned to a confidence rating condition, eighty participants were included in the analysis. These results are presented in Figure 6.

As in the analysis of accuracy, subjective confidence was affected by the interaction between Categorization Rule and Experimental Block, $F(11,858) = 3.32$, $MSE = 69.37$, $p = .011$, $\eta^2_p = .04$. A decomposition of this effect revealed significant simple effects in both the rule-based, $F(11,858) = 11.69$, $MSE = 28.82$, $p < .001$, and the information-integration conditions, $F(11,858) = 17.90$, $MSE = 28.82$, $p < .001$. Participants expressed greater certainty in their responses when learning a rule-based category structures earlier in the course of the experiment in comparison to those participants who learned an information-integration category structure (see Figure 6).

In the same manner as the accuracy analysis, a significant effect of Experimental Block was also observed, $F(11,858) = 36.81$, $MSE = 69.37$, $p < .001$, $\eta^2_p = .321$, as well as a marginal effect of Categorization Rule, $F(1,78) = 2.97$, $MSE = 720.38$, $p = .089$, $\eta^2_p = .037$. In general, participants' confidence increased across experimental blocks with greater confidence being expressed in the rule-based condition relative to the information-integration condition (see Table 2).

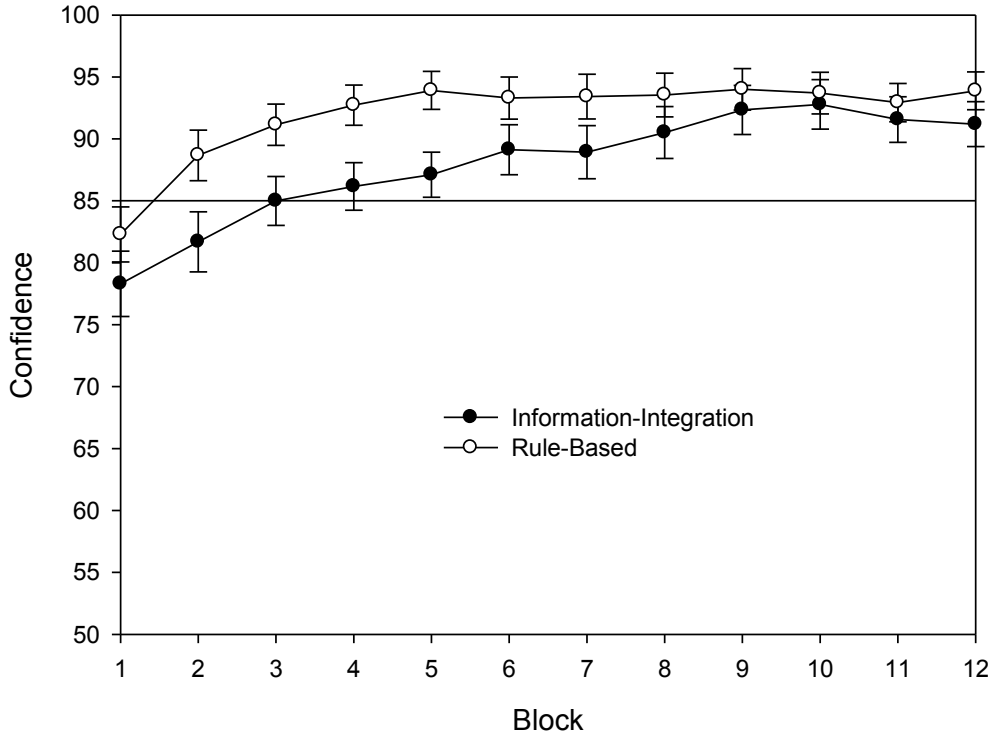

Figure 6. Mean trial-level confidence across category rule conditions. A reference line indicates the 85% performance asymptote. The error bars represent the standard error of the mean.

4.2.3.2 Trial-Level Subjective Confidence Calibration. As described in the Appendix, calibration was computed by taking the proportion correct for a given confidence category, weighting it by the number of observations within this category and taking the sum of the deviations from perfect calibration. Due to inter-block variability resulting from individual differences in confidence reports, the dataset was aggregated into experimental phases in order to compare training and transfer phases prior to conducting this analysis. Given the correspondence between the general trends for accuracy and confidence analysis, this transformation was conducted to reduce increased variability within subjective confidence reports. In the following analysis, Blocks 1 and 2 are averaged into Phase 1 (Early Training), Blocks 3 through 6 are averaged into Phase 2 (Intermediate Training), Blocks 7 through 10 are averaged into Phase 3 (Late Training), and the transfer Blocks 11 and 12 are averaged into Phase 4 (Transfer). This transformation allows for a direct comparison of transfer phase where performance was comparable across blocks to early and late stages of training where greater differences in calibration were likely to be observed. The subsequent design consisted of a 2 (Categorization Rule: rule-based vs. information-integration) x 4 (Experimental Phase: 1-4). Of critical importance, an examination of trial-level subjective calibration obtained significant improvements in calibration across Experimental Phases (See Table 3), $F(3,258) = 9.57$, $MSE = .001$, $p < .001$, $\eta^2_p = .10$, as well as a marginal effect of Categorization Rule (see Table 2), $F(1,86) = 3.83$, $p = .054$, $\eta^2_p = .04$, whereas their interaction was not significant, $F(3,258) = .18$, $p = .87$. Observing significant improvements in subjective calibration could be interpreted as indicative of improvements to participants' assessments of their performance. Moreover, the marginally significant effect of categorization rule might additionally suggests that participants' assessments of their performance exhibited greater bias in the rule-based condition relative to the information-integration condition. It is

clear, however, that increases in categorization accuracy over experimental phases might be the cause of improved calibration rather than a reduction in subjective confidence level.

Table 3. Mean subjective calibration indices for trial-level confidence. Standard errors are in parentheses.

Ex.	Cond.	Early	Intermediate	Late	Transfer
1	RB	.03 (.00)	.02 (.00)	.02 (.00)	.02 (.00)
	II	.04 (.00)	.03 (.00)	.03 (.00)	.03 (.00)
2	RB	.03 (.00)	.02 (.00)	.02 (.00)	.02 (.00)
	II	.03 (.00)	.02 (.00)	.02 (.00)	.02 (.00)
3	RB	.03 (.01)	.02 (.00)	.02 (.00)	.03 (.00)
	II	.04 (.01)	.03 (.00)	.03 (.00)	.03 (.00)
4	RB	.07 (.01)	.06 (.01)	.06 (.01)	.06 (.01)
	II	.06 (.01)	.04 (.01)	.05 (.02)	.04 (.01)

4.3.2.3 Trial-Level Overconfidence Bias.

Overconfidence was computed by obtaining the mean difference between mean confidence and mean accuracy for each condition (see Appendix). The mixed-design ANOVA was identical to that used to analyze mean trial-level confidence and calibration: 2 (Categorization Rule: rule-based vs. information-integration) x 4 (Experimental Phase: 1-4). It revealed an interaction between Experimental Phase and Categorization Rule, $F(3,234) = 3.06$, $MSE = .01$, $p = .038$, $\eta^2_p = .038$. A decomposition of this effect revealed significant simple effects for both rule-based, $F(3,234) = 5.45$, $MSE = .01$, $p = .001$, and information-integration

category structures, $F(3,234) = 7.53$, $MSE = .01$, $p < .001$. In contrast, I did not observe a significant effect of Experimental Phase alone, $F(3,234) = .92$, $p = .34$, nor Categorization Rule, $F(1,78) = .03$, $p = .863$.

As can be seen in Figure 7, overconfidence remained relatively constant in both conditions after the initial phase of training. In the information-integration condition, I observed a reduction in overconfidence suggesting that participants' subjective confidence decreases after the first two blocks of trials in the Early Training Phase. Such a pattern could have resulted from the receipt of block-level feedback producing a more conservative criterion for confidence responses. On such an account, whereas participants receiving block-level feedback might predict that their performance would continue to improve, in the absence of block-level feedback they might not overestimate their performance on subsequent blocks of trials. In contrast, an increase in overconfidence was observed in intermediate phases of training in the rule-based condition. This finding might suggest that once participants identified the one-dimensional rule, they expected to have continual improvements in performance. Such a result would be expected if participants ignored negative feedback for exemplars that represent exceptions to a category boundary.

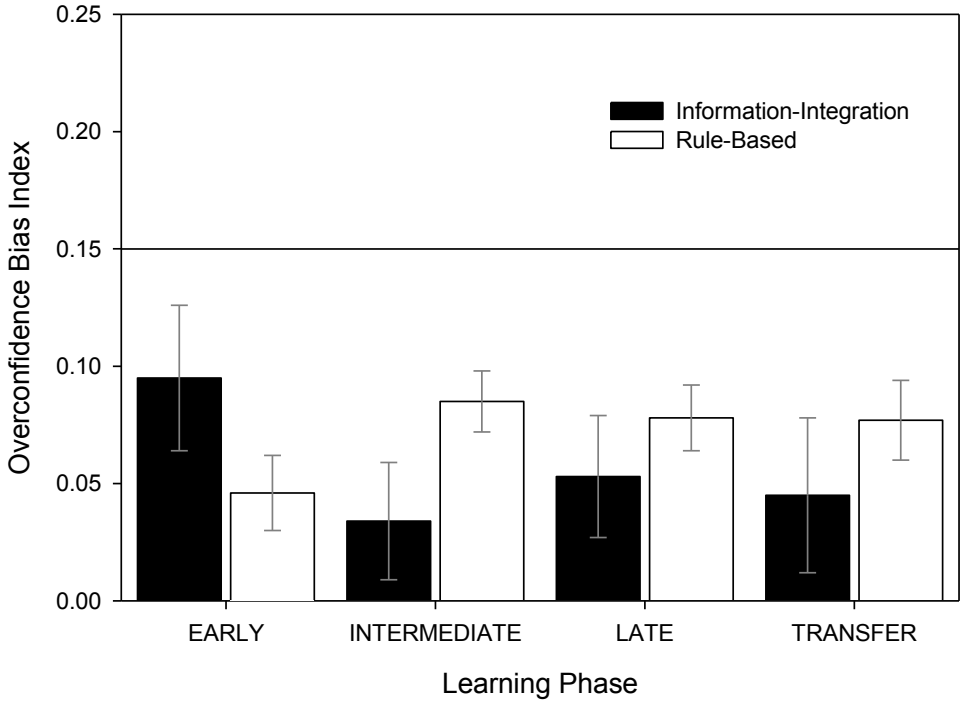

Figure 7. Average overconfidence across category rule conditions between experimental phases for trial-by-trial confidence. A reference line at 0.15 indicates the expected degree of overconfidence if participants reach the performance asymptote while expressing 100% confidence in their responses. The error bars represent the standard error of the mean.

4.2.3 Trial-Level Confidence Response Time.

A mixed-design ANOVA was conducted on confidence response times (CRT) with using a design identical to the one used for mean confidence. Experimental block significantly affected CRT, $F(11,858) = 164.649$, $MSE = 21021$, $p < .001$, $\eta^2_p = .679$. Neither categorization condition, $F(1,78) = .09$, $p = .76$, nor its interaction with experimental block were significant, $F(11,858) = .95$, $p = .44$, however. As Figure 5 demonstrates, participants' CRT became faster across experimental block. Hence, DRT and CRT learning curves seemed to follow a similar pattern

across training, but CRTs were much faster overall. This replicates the finding that, in the absence of the requirement of speeded responses, post-decisional confidence reports might be computed along with the primary decision process (e.g., Baranski & Petrusic, 1998).

4.3 Discussion

Experiment 1 replicated the findings of previous categorization studies. Participants required more trials to learn an information-integration category structure relative to those who learned a rule-based category structure. In general, a decrease in DRT was also observed across blocks of trials suggesting that participants' response selection became increasingly automatic with training. Given that these categorization results were comparable to other studies (e.g., Ell & Ashby, 2006), the results of confidence reports can be examined and interpreted without qualification.

In terms of categorization accuracy, the requirement of confidence reports had neither a facilitative nor detrimental effect on participants' ability to learn the category structures. No differences in response accuracy were observed in either the information-integration or the rule-based condition when confidence was required. In contrast, response latencies increased for both category structures when confidence reports were required. A direct means to account for the present finding is provided by the confidence processing literature. For instance, Baranski and Petrusic (2001) observed that increased response latencies that result from the requirement of confidence decreases as a function of experimental block. They suggested that such a pattern would be expected if the process of mapping subjective probability of being correct onto a confidence response category became increasingly automatic. This suggests that confidence processing is initially an effortful process independent of the requirements of categorization. The

observation that categorization performance did not change with the requirement of confidence reports suggests that the confidence process is sampling the same source of information as the categorization decision.

In contrast to single-process models of confidence processing (e.g., Balakrishnan & Ratcliff, 1996; Ferrel & McGooey, 1980), the finding of decreased response latencies for confidence responses suggests that confidence reports required an additional, secondary process that can itself be automated. The results of the present analyses also suggest that whereas a similar set of rescaling operations is used during the generation of a confidence report, the source of information used in the rule-based and information-integration conditions differed. Greater trial-level overconfidence in the rule-based condition suggests that the greater accessibility of the stimulus representations produced greater certainty in participants' responses than was warranted.

In conclusion, the calibration indices that were measured in the present experiment are informative with regards to the nature of the representations that the participants used to classify the stimuli. In general, participants became less overconfident on a trial-by-trial basis in the information-integration condition as training progressed whereas participants in the rule-based condition became more overconfident. As is clear from Figure 4, participants in the rule-based condition reached the performance asymptote in the intermediate phase of training. The observed overconfidence would therefore be a result of a proportionally greater increase in response certainty relative to response accuracy. An increase in confidence would be observed if participants predicted that their performance was greater than objectively justified by the proportion of negative feedback that they received. Thus, in the absence of information provided to participants resulting from feedback, the basis for this prediction appears to be the

representation of the category structure. At a phenomenological level, participants must be experiencing an accessible representation that has the appearance of accuracy. Decreases in overconfidence in the information-integration condition could be explained by a greater reliance on negative feedback resulting from the engagement of the procedural-learning system. Although this could be the result of either increases in the accuracy of representation used to categorize stimuli or that used to inform confidence (or the conjoint influence of both of these factors), a comparison of mean confidence and overconfidence bias suggests the origin of this effect. Observing these difference between confidence and accuracy in the rule-based and information-integration conditions suggests that two representations are available to participants: one that is accessible to participants in the rule-based condition and one that is less accessible to participants in the information-integration condition.

Before assuming that these results can be used to infer a distinction between explicit and implicit representations and their respective learning systems, another methodological manipulation must be considered. In order to replicate the methods used in typical categorization experiments, Experiment 1 provided block-level feedback to participants. Block-level feedback rather than trial-level feedback or differences in representation accessibility could have led participants in the information-integration condition to report less confidence in their responses in subsequent blocks of trials. As a consequence, the difference in overconfidence between the conditions might have been caused by the block-level feedback rather than the accessibility of the category structure or reliance on trial-level feedback.

5.0 Experiment 2

The results of Experiment 1 appear to provide evidence for a dissociation between the representations used within the explicit system when participants learn rule-based category structures and those used within the implicit system when participants learn information-integration category structures. The effect of block feedback, however, might have reduced overconfidence bias in the information-integration condition. On this interpretation, after receiving block feedback, participants in the information-integration condition might have become more conservative in their confidence ratings. If such behaviour did occur, it would suggest that the participants had an awareness of the category structure available within their hypothesis-testing system but began to regard it as nondiagnostic of the category structure. Such a relationship conforms to a two-process account of categorization. Alternatively, rather than two categorization systems, the two processes that give rise to the differences between conditions might reflect the operations used to categorize stimuli and generate confidence reports. On this interpretation, a single system encodes stimuli as being members of two categories. Following the receipt of feedback, stimulus information, long-term memory representations of category members, and non-diagnostic information would be available to be used by an additional confidence process. On this account, confidence reports would be a function of information that does not necessarily correspond to the categorization process. Given the reasonable degree of correlation between accuracy and confidence reports, however, confidence reports appear to be primarily determined by information pertaining to the categorization process.

In Experiment 2, I examined whether a greater dissociation between explicit and implicit measures could be obtained if participants were not made aware of their overall performance for each block of trials. In general, the pattern of response accuracy, DRT, mean confidence, and

CRT should be comparable to Experiment 1. Without block feedback, however, I predicted that a greater overconfidence bias should be evidenced through early stages of learning given that participants' confidence reports should be primary based on explicit representations.

5.1 Method

5.1.1 Participants and Design.

One-hundred and seventeen participants were recruited from Carleton University and were awarded course credit for their participants in a single 2-hour session. Participants were required to learn a rule-based category structure (defined by frequency; n = 58) or an information integration category structure (defined by frequency and line orientation; n = 59). Half of the participants were assigned to the block-confidence condition (n = 55) whereas the other half were assigned to the trial-and-block confidence condition (n = 62).

5.1.2 Stimuli and Procedure.

Stimuli were identical to those used in Experiment 1. The design of the study was also identical to that of Experiment 1 except that no block-level feedback was provided.

5.2.0 Results

Using the same exclusion criterion of Experiment 1, participants who obtained less than 65% correct on training trials were excluded from further analysis. This procedure resulted in the exclusion of (n = 4) from the rule-based condition and (n = 3) from the information-integration condition. As in Experiment 1, there were no significant differences for the learning criterion in comparison to the final learning block (M = .8534, SD = .073), $t(118)$ = .50, p = .617, or the

mean for the last two learning blocks (M = .8547, SD = .068), $t(118) = .75$, $p = .458$. Moreover, comparisons of guessing responses (50%) to performance across all blocks as well as learning blocks and transfer blocks revealed that participants had in fact learned to differentiate features of the responses categories above chance, all ts > .62, all ps < .001.

5.2.1 Proportion Correct.

The experimental design was identical to that of Experiment 1. A 2 (Categorization Rule: rule-based vs. information-integration) x 2 (Confidence Condition: block-level confidence vs. trial-and-block confidence) x 12 (Experimental Blocks: 1-12) mixed-design ANOVA was performed on proportion correct, collapsing across response key. First, a main effect of Experimental Blocks was observed in categorization response accuracy, $F(11,1243) = 43.33$, $MSE = .006$, $p < .001$, $\eta^2_p = .28$. Experimental Block and Categorization Rule also had a significant interaction, $F(11,1243) = 3.02$, $MSE = .006$, $p = .012$, $\eta^2_p = .03$. A decomposition of the ANOVA revealed significant simple effects in both the rule-based, $F(11,1265) = 12.66$, $MSE = .01$, $p < .001$, and information-integration conditions, $F(11,1265) = 33.56$, $MSE = .01$, $p < .001$. As is clear from Figure 8, participants who learned rule-based category structures reached asymptote faster than those who learned information-integration category structures. These findings again conform to dual-process accounts of categorization. Neither the main effects of Categorization Rule, $F(1, 113) = 1.68$, $p = .20$, and Confidence Condition, $F(11,113) = .147$, $p = .70$, nor their interaction, $F(11,113) = .09$, $p = .76$, reached significance, however.

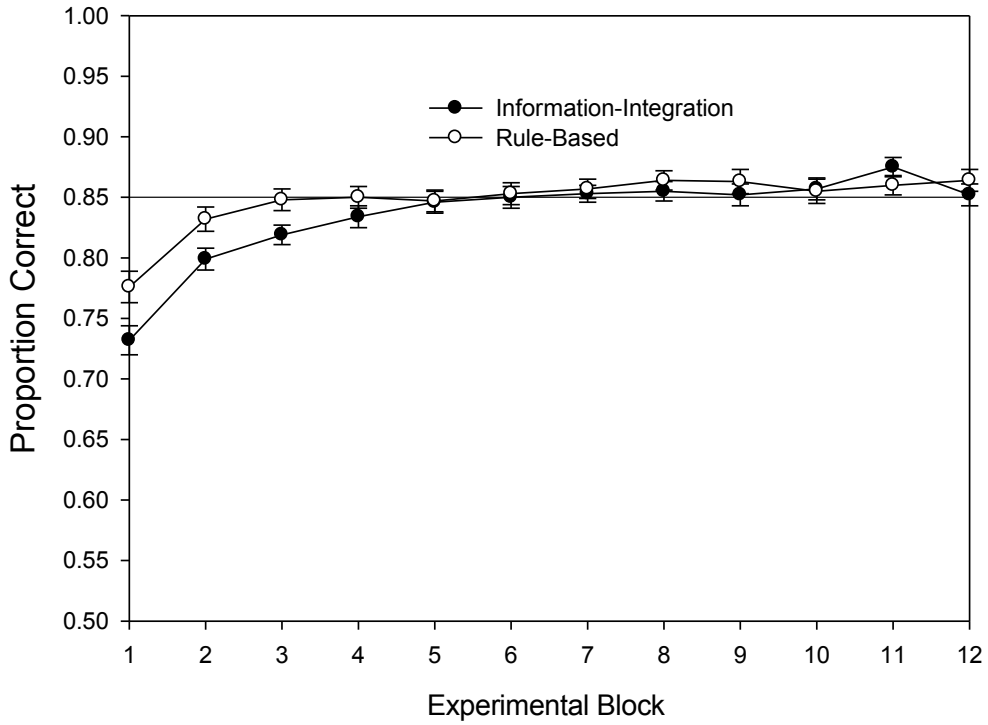

Figure 8. Categorization response accuracy for Experiment 2. The error bars represent the standard error of the mean.

5.2.2 Decision Response Time.

Decision response time was analyzed using the same design as response accuracy. The DRT results obtained in the present experiment were more complex than those observed in Experiment 1. Replicating the results of Experiment 1, I observed significant main effects of Experimental Block, $F(11,1243) = 17.51$, $MSE = 313240$, $p < .001$, $\eta^2_p = .13$, Categorization Rule, $F(1, 113) = 3.87$, $MSE = 1677380$, $p = .052$, $\eta^2_p = .03$, as well as their interaction, $F(11, 1243) = 2.85$, $MSE = 313240$, $p = .012$, $\eta^2_p = .03$. Figure 9 provides mean performance in this task. A decomposition of the ANOVA revealed significant simple effects in both the rule-based, $F(11,1265) = 5.58$, $MSE = 161857$, $p < .001$, and information-integration conditions, $F(11,1265)$

= 13.78, *MSE* = 161857, $p < .001$. These results suggest that participants' categorization process differed across experimental blocks for the categorization rules.

As in Experiment 1, I sought to determine whether the requirement of confidence affected the time taken to categorize stimuli. Confidence Condition was also found to have a significant effect alone, $F(1,113) = 23.16$, $MSE = 1677380$, $p < .001$, $\eta^2_p = .20$, and in combination with Categorization Rule, $F(1,113) = 11.84$, $MSE = 1677380$, $p = .001$, $\eta^2_p = .10$, and in a three-way interaction with experimental block, $F(11, 1243) = 2.84$, $MSE = 313240$, $p = .013$, $\eta^2_p = .03$. This pattern again indicates that the requirement of confidence reports increased primary decision response time, suggesting that the confidence processing occurred concurrently with the primary decision. Moreover, response times were longer for the information-integration condition with the requirement of confidence suggesting that the provision of confidence reports required processes other than the procedural-learning categorization system used for response selection.

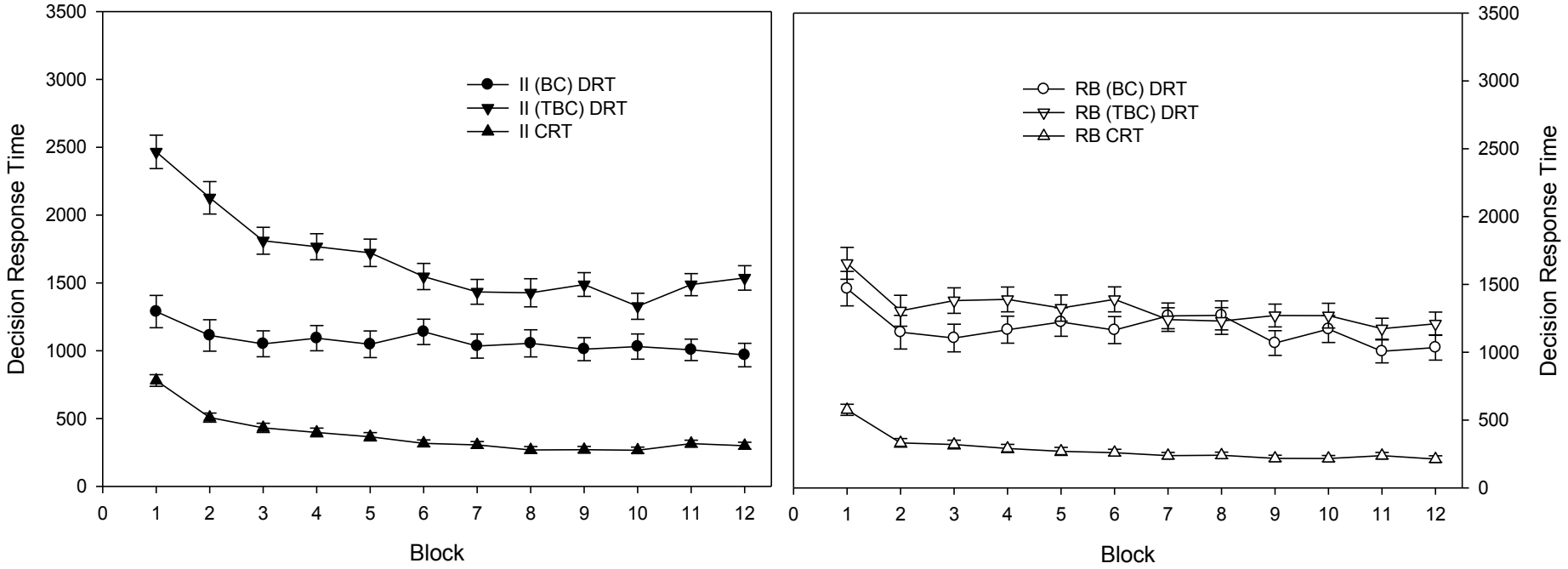

Figure 9. Categorization decision response time (DRT) for trial- and block-level (TBC) and block-level only (BC) confidence reports, and confidence response time (CRT) over experimental blocks. The error bars represent the standard error of the mean.

5.2.3 Confidence

5.2.3.1 Mean Trial-Level Confidence. Confidence reports were analyzed in the same manner as Experiment 1, with n = 62 in the confidence condition. An analysis of mean confidence revealed a significant effect of Experimental Block, $F(11,660) = 37.35$, $MSE = 52.27$, $p < .001$, $\eta^2_p = .38$. This general pattern again suggests that participants had a general awareness that their performance was increasing from one block to the next. Unlike Experiment 1, I failed to obtain a main effect of Categorization Rule, $F(11,660) = 1.42$, $p = .24$, or its interaction with Experimental Block, $F(11,660) = 1.16$, $p = .33$. As is clear from an inspection of Figure 10, mean confidence rapidly increased from Block 1 to Block 2 but only increased moderately thereafter. This result suggests that participants reported an invariant level of confidence over categorization rules.

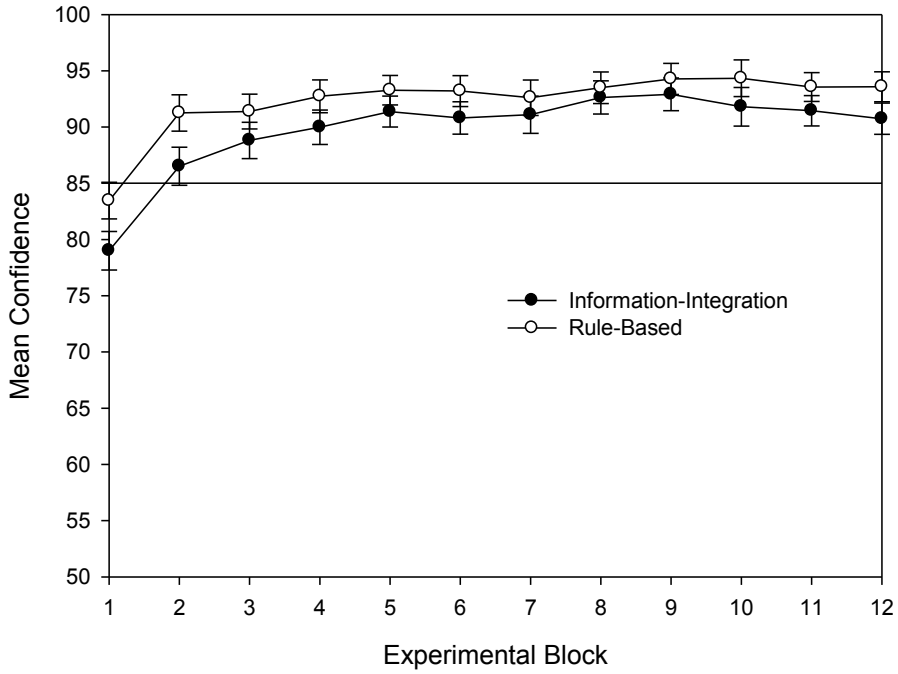

Figure 10. Mean trial-level confidence across category rule conditions. A reference line indicates the 85% performance asymptote. The error bars represent the standard error of the mean.

5.2.3.2 Trial-Level Subjective Confidence Calibration. To qualify the results of the confidence analysis, subjective calibration was again examined. Hence, a 2 (Categorization Rule: rule-based vs. information-integration) x 4 (Experimental Phase: Early, Intermediate, Late, and Transfer) repeated-measures ANOVA was conducted. Replicating the results of Experiment 1, subjective calibration improved over Experimental Phase (see Table 3), $F(3,180) = 13.82$, $MSE = 103.25$, $p < .001$, $\eta^2_p = .19$. As in Experiment 1, these results could indicate that participants became better judges of their performance with more experimental blocks. Alternatively, it could indicate that participants' performance improved over experimental blocks while initially high estimates of their performance remained constant. When considered in conjunction with the accuracy and confidence results provided above, this latter conclusion appears to be well supported. Namely, mean confidence approached an asymptote shortly after Block 2 whereas accuracy continued to grow in both rule-based and information-integration conditions.

Again, I failed to obtain either a significant main effect of Categorization Rule (see Table 2), $F(1,60) = .20$, $p = .657$, or its interaction with Experimental Phase, $F(3,180) = .53$, $p = .56$. This finding suggests that participants' ability to judge their performance in a task might be independent of the representation of the response category on a trial-to-trial basis. Again, this could be taken as evidence for an independence of confidence processing and categorization systems.

5.2.3.3 Trial-Level Overconfidence Bias. An analysis of overconfidence revealed the same general trend as the calibration analysis, with a significant effect of Experimental Phase, $F(3,180) = 3.13$, $MSE = 151.70$, $p = .035$, $\eta^2_p = .05$. As Figure 11 demonstrates, overconfidence was again observed in the present experiment. In general, participants' overconfidence increased

over the course of experimental phase. Although not significant, greater overconfidence was again observed in the rule-based condition relative to the information-integration categorization condition (see Table 2), $F(1,60) = .68$, $p = .41$. Moreover, greater overconfidence was also observed in the rule-based condition in the final experimental phase although this interaction was not significant, $F(3,180) = .14$, $p = .91$. Along with other studies examining the effect of feedback on confidence, Experiments 1 and 2 suggest that trial-level feedback affects accuracy and confidence reports in a different fashion.

Prior to accepting such an account, it is critical to note the fairly important difference in the magnitude of the overconfidence bias found in Experiment 2 relative to Experiment 1 when the early phases of training are considered. Namely, if block feedback reduced overconfidence bias in the information-integration condition, then I should still have observed high overconfidence in early phases of training as well as a continuation of this trend throughout subsequent blocks of trials. One possibility is that the absence of greater overconfidence could simply represent individual differences between groups of participants in Experiments 1 and 2. If this were the case, participants in Experiment 1 could have been predisposed to report higher confidence, but when confronted with feedback indicating that they were incorrect, they might have become more conservative in their responses. In contrast, participants in Experiment 2 might have been less inclined to overconfidence.

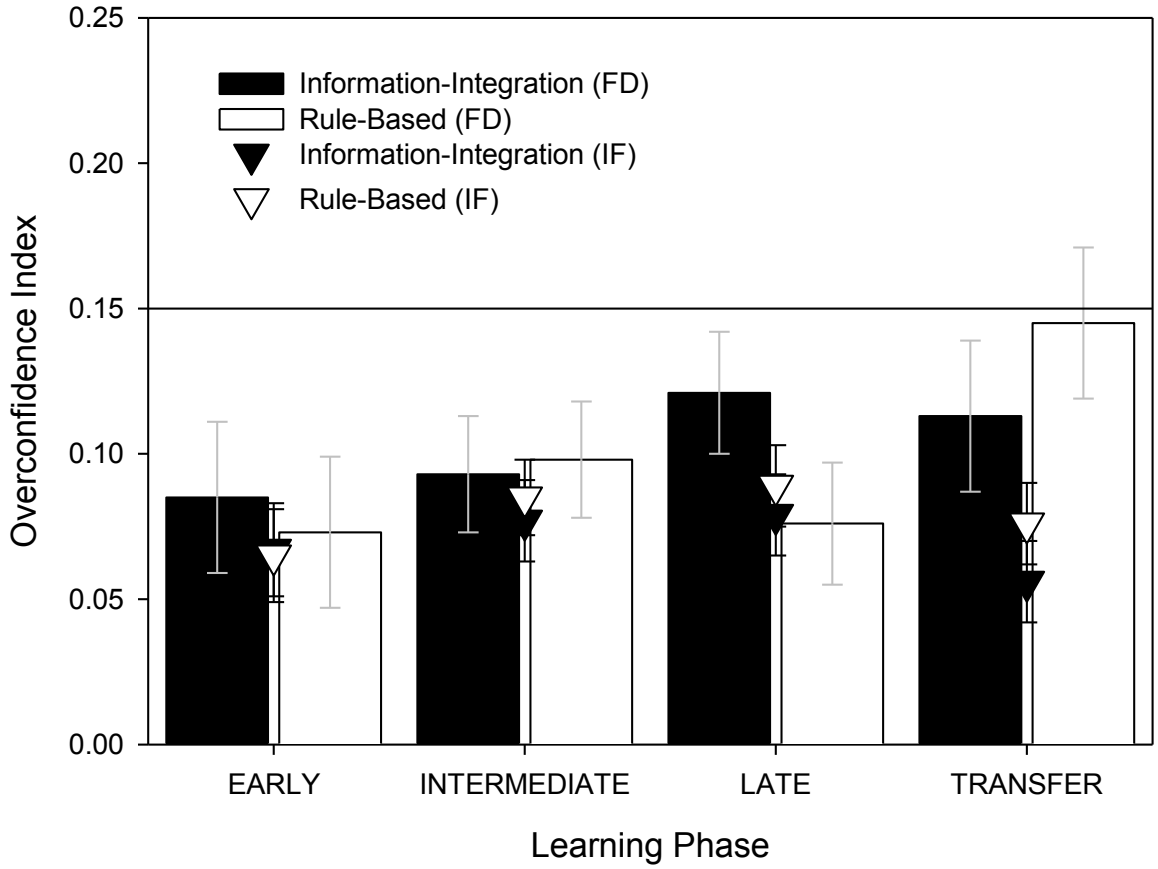

Figure 11. Overconfidence bias for Experiments 2 (Triangles) and Experiment 3 (Bars) for immediate feedback in Experiment 2 (IF) and delayed feedback in Experiment 3 (DF). Reference line represents maximum overconfidence (.15). The error bars represent the standard error of the mean.

The importance of individual difference has been observed in several studies (e.g., Pallier, Wilkinson, Danthiir, & Kleitman, 2002; Stanovich & West, 1998), but I will consider an alternative explanation given the results of the accuracy analysis in comparison to Experiment 1. When we consider the obtained proportion correct was considered, it was found that there was less of a difference between the rule-based and information-integration conditions in early phases

of training. Increased accuracy in the information-integration condition would lead to less overconfidence given comparable levels of accuracy and confidence. Thus, in the absence of evidence for individual differences in responses certainty, the difference in this effect appears to result from differences in participants' ability to acquire the category structure. Thus, a further manipulation is required to provide more concrete support for the account of categorization learning systems advanced here. This will constitute the method of Experiment 3 wherein I delayed feedback in order to reduce learning in the information-integration condition while leaving the rule-base condition unaffected.

5.2.4 Trial-Level Confidence Response Time.

A mixed-design ANOVA was conducted on confidence response times (CRT) in the same manner as Experiment 1. The results are shown in Figure 9. Supporting DRT results, there was a main effect of Experimental Block, $F(11,660) = 118.38$, $MSE = 22431.91$, $p < .001$, $\eta^2_p = .66$, as well as its interaction with Categorization Rule, $F(11,660) = 5.14$, $MSE = 22431.91$, $p = .001$, $\eta^2_p = .08$. A decomposition of this interaction revealed significant simple effects for both the rule-based, $F(11,660) = 42.45$, $MSE = 7756$, $p < .001$, and information-integration conditions, $F(11,660) = 78.74$, $MSE = 7756$, $p < .001$. As with the DRT, participants response latency decreased across experimental blocks with participants in the information-integration condition responds more slowly than participants in the rule-based condition. Given the similar pattern in DRT, this suggests that confidence processing occurs during both the categorization response selection as well as post-decisionally supporting previous studies within the confidence processing literature (Baranski & Petrusic, 1998).

5.3 Discussion

The results of Experiment 2 generally conformed to those obtained in Experiment 1. Replicating results of previous studies (e.g., Ashby et al. 2002), I found that participants required more trials to learn the information-integration category structure in comparison to rule-based category structures. With sufficient training, however, participants in both rule-based and information-integration conditions performed equally well. Similarly, I was also able to demonstrate that decision response times decreased across experimental blocks. This suggests that response selection processes became progressively more automated with greater training. Concomitantly, the requirement of confidence processing slowed categorization responses made by the primary decision process suggesting that confidence reports were processed concurrently. This pattern was more pronounced in the information-integration condition where both participants' DRTs and CRTs were longer than in the rule-based condition. This provides evidence for two representations. When confidence is required, if the representation within the hypothesis-testing is similar to that used to report confidence, I would expect shorter DRTs in the rule-based condition relative to the information-integration condition. The longer DRTs observed within the information-integration condition would be indicative of the need to re-scale the evidence accumulated by a procedural-learning system into an explicit representation that could be used to report confidence. The implications of the obtained results are twofold. First, the hypothesis-testing system and confidence processing appear to use similar representations. Little additional processing is required to change representations from that used to categorize stimuli in the rule-based condition relative to the information-integration condition. In contrast, the procedural-learning system requires more time to process confidence *concurrently* with the categorization response. This suggests that two different representations are required: one

representation to successfully complete response selection and another to report confidence. Alternatively, others have argued that the primary difference between rule-based and information-integration category structures is that the former are easier to learn (e.g., Newell, et al. 2010; Nosofsky & Johansen, 2000). Even if this were the case, an interaction of categorization structure and the requirement of confidence report suggest that difficulty has separate, non-additive effects and is consistent with multiple processing stages. These stages could reflect either the re-scaling of evidence from a procedural-learning system to confidence report or a central bottleneck for response selection of a concurrent confidence processes. In either case, two processes are occurring further suggesting some qualitative difference in their respective representations.

Similar patterns to those obtained in Experiment 1 were also evidenced in confidence responses. In general, participants' confidence quickly reached an asymptote and continued to grow across experimental blocks. Together with increases in categorization response accuracy, this also gave rise to improvements in calibration across experimental blocks. On these grounds, these gains in calibration could be considered artifactual in that participants only became better calibrated due to increases in accuracy. Such a conclusion is supported by overconfidence results: although participants became better calibrated across experimental trials, they expressed greater certainty than was warranted by their performance.

The results of Experiment 2 provide additional confirmation for a dissociation between a hypothesis-testing system that uses explicit representations and a procedural-learning system that uses implicit representations. This was evidenced by additional categorization response time in the information-integration condition. The pattern of differences in overconfidence between the rule-based and information-integration conditions was also suggestive, but the failure to obtain

significant results present at least two possibilities. First, the greater overconfidence observed in the rule-based condition in Experiment 1 might have been caused by block-level feedback altering subjective confidence. On this account, participants' confidence reports would be based on an accessible, explicit representation in both the rule-based and information-integration conditions. Although participants in the rule-based condition would receive block-level feedback that reinforces their high confidence, participants in the information-integration condition would receive feedback suggesting that they are performing below their subjective level of confidence. This latter group of participants would conceivably increase the criterion amount of evidenced required to report such high levels of subject certainty, thereby reducing their confidence. Such an account assumes that even if categorization systems are dissociable, confidence reports sample a different representation and cannot be used to assess the representations within each categorization system.

An alternative explanation is also possible. It might be the case that the explicit representation within the hypothesis-testing system could be used to report confidence, but some individual differences might have obscured this result. As I noted in the Results section of this experiment, individual differences could either be evidenced in subjective certainty (e.g., Pallier et al., 2002) or the ease with which participants learned the category structures. Relative to Experiment 1, participants in Experiment 2 had higher levels of accuracy in early training blocks. Differences in overconfidence would not be observed under these conditions do to a high degree of correspondence between accuracy and confidence.

Two possible sources of differences between categorization and confidence processes appear plausible based on the extent to which categorization is sampled by the confidence process. First, differences in accuracy and confidence reports would result from the accessibility

of a representation evidenced as a constant proportional bias (e.g., 15% greater than accuracy). In this case, the more accessible a representation is to participants, the greater the bias in assessing their performance. A constant proportional bias could be seen as a function of scaling evidence from one format (e.g., perceptual properties) to another (e.g., a proportion of accumulated evidence). On this account, confidence reports would be based on re-scaled evidenced accumulated during stimulus discrimination and categorization. Alternatively, a constant level of confidence would be expected (e.g., 95% confidence in responses) if subjective reports had little to no relation with performance in a task (e.g., see, Lichtenstein & Fischhoff, 1977 Experiments 1 and 2; Nisbet & Wilson, 1977; Wells, Lindsay, & Ferguson, 1979). A constant confidence level would occur if the availability of an explicit representation produced high confidence regardless of its correspondence to the stimulus representation. On this account, confidence reports would be independent of categorization responses, suggesting that categorization and confidence processes constitute separate sets of operations.

In two subsequent experiments, I examine the effect of trial-level feedback on the respective categorization systems by delaying feedback (Experiment 3) and by increasing the proportion of negative feedback received by participants (Experiment 4). According to previous studies (Maddox et al., 2003), delayed trial-level feedback should affect only the procedural-learning system creating differences in response accuracy. Under these conditions, a constant level of confidence would therefore create considerably greater overconfidence in the information-integration condition due to the reduction in accuracy. If instead feedback is used to modify the criterion used to assess certainty in one's performance, greater overconfidence should be observed in the rule-based condition when feedback is removed or delayed by 2500 ms. In contrast, an increase in the proportion of negative feedback resulting from a change in

performance asymptote should show differences in confidence reports due to differential accessibility of representations used within the respective categorization systems. When performance is heavily constrained (e.g., accuracy cannot exceed 65%), participants in both rule-based and information-integration conditions would quickly arrive at a performance asymptote. A constant level of confidence bias should therefore produce equivalent overconfidence in each condition. A proportional increase in overconfidence based on accessibility would instead result in greater overconfidence when representations used to categorize performance are explicit.

6.0 Experiment 3

One method used in previous implementations of the randomization technique to produce greater differences in the rule-based and information-integration conditions is the introduction of a feedback delay (e.g., Maddox et al., 2003). In this version of the task, participants categorize the stimuli but, prior to receiving feedback, they receive an unfilled inter-stimulus interval in which no stimuli or information is presented. In these conditions, performance is reduced within the information-integration whereas it remains unaffected in the rule-based condition (Maddox et al., 2003). Thus, even if individual differences in accuracy and confidence ratings occur, the provision of a feedback delay should result in decreased accuracy within the information-integration condition only.

A feedback delay allows for direct tests of the assumptions concerning representations and confidence processing. If confidence reports represent a proportionate bias that results from rescaling primary decision information, a reduction in mean confidence should be observed in relation to the reduction in categorization accuracy. Overall, this would lead to similar levels of overconfidence in both rule-based and information-integration conditions (e.g., 10-15%). Such a

finding would suggest that participants have equivalent access to the representations used in both categorization conditions. This would provide evidence against dual-process accounts of categorization such as COVIS.

Alternatively, a dual-process account of categorization would be supported if differences were obtained within accuracy and confidence between rule-based and information-integration conditions. Subjective calibration should be worse in the information-integration condition for two reasons. First, in line with previous studies of delayed feedback (e.g., Maddox et al., 2003), I would expect reduced accuracy due to a failure to adequately acquire an accurate representation within the procedural-learning system. Concurrently, I would still expect a relatively constant level of confidence resulting from the continued accessibility of the explicit representation used to report confidence. In contrast to accounts of single source of information, both explicit representations and negative feedback could conjointly influence confidence processing. If confidence reports are affected by representations within the explicit system as well as response feedback, then a feedback delay should differentially affect rule-based and information-integration category structures over the course of learning. If the accessibility of explicit representations is the primary determinant of confidence reports, overconfidence bias should remain relatively constant across training blocks. In the transfer phase, we might expect greater overconfidence due to a removal of negative feedback. In this way, participants would lack an additional source of information leading to less conservative responding.

6.1 Method

6.1.1 Participants and Design.

Sixty participants were recruited from Carleton University and were awarded course credit for their participants in a 3-hour session. Participants were randomly assigned to one of two experimental conditions. They were required to learn a rule-based category structure (defined by frequency; n = 29) or an information integration category structure (defined by frequency and line orientation; n = 31). All participants reported confidence and received delayed feedback.

6.1.2 Materials and Procedure. The materials were identical to those outlined in the General Methods Section. The procedure followed the one outlined in Experiment 2 with the exception of one critical feature. Following their categorization decision and confidence reports, participants were required to wait 2500 ms prior to receiving feedback. Due to the requirement of an intervening confidence report between categorization response and feedback, the feedback delay was variable. Thus, the delay is given by the sum of CRT (i.e., the interval between categorizing a stimuli and providing a confidence report) and the 2500 ms feedback delay. Importantly, however, previously studies have established that 2500 ms is the minimum time required to reduce learning in information-integration conditions (e.g., Maddox et al., 2004). Thus, longer delays have the potential to reduce performance to an even greater extent. Thus, this variability would increase the difference between the II and RB conditions. Unlike the previous Experiments 1 and 2, only a confidence condition was used. In every other respect, feedback was provided in the same manner as above. No block-level feedback was provided to participants.

6.2 Results

In the same manner as Experiments 1 and 2, participants' accuracy was analyzed in order to establish that they performed above chance and reached the 85% learning criterion. Prior to this analysis, participants failing to obtain 65% correct over all training blocks were eliminated from the analysis. This resulted in the exclusion of no participants from the rule-based condition (n = 0) whereas some were eliminated from the information-integration (n = 4) conditions. This result follows from the nature of the feedback delay paradigm. Mean performance for both the final training block (M = .84, SD = .08) and the last two training blocks (M = .84, SD = .07) fell somewhat below the learning criterion. This difference was found not to be significant in either a comparison with the learning criterion and both the final training block, $t(62) = -1.35$, $p = .183$, or the last two training blocks, $t(62) = -1.31$, $p = .196$.

6.2.1 Proportion Correct.

The experimental design of Experiment 3 represents a minor modification of the design used for Experiments 1 and 2. A 2 (Feedback Delay: no-delay vs. 2500* ms delay) x 2 (Categorization Rule: rule-based vs. information-integration) x 12 (Experimental Blocks: 1-12) mixed-design ANOVA was performed on proportion correct, collapsing across response key. Categorization accuracy was affected by both Experimental Block, $F(11,638) = 14.92$, $MSE = .01$, $p < .001$, $\eta^2_p = .205$, as well as Categorization Rule (see Table 2), $F(1,58) = 8.47$, $MSE = .04$, $p = .005$, $\eta^2_p = .13$. Their interaction was not significant, $F(11,638) = .527$, $p = .78$, however. These results are presented in Figure 12. As in previous experiments, participants' categorization accuracy increased over the experimental blocks with performance also being affected by the category structure provided to participants. Participants learning the rule-based category

structure were more accurate than then those learning the information-integration category structure. The latter participants also failed to reach the performance asymptote. When compared to performance in Experiment 2, these results replicate findings of performance decrements with feedback delay (e.g., Maddox et al., 2003). Delayed feedback reduced categorization accuracy for participants assigned to the information-integration whereas performance in the rule-based condition was equivalent to the results obtained in Experiments 1 and 2 (Figure 12).

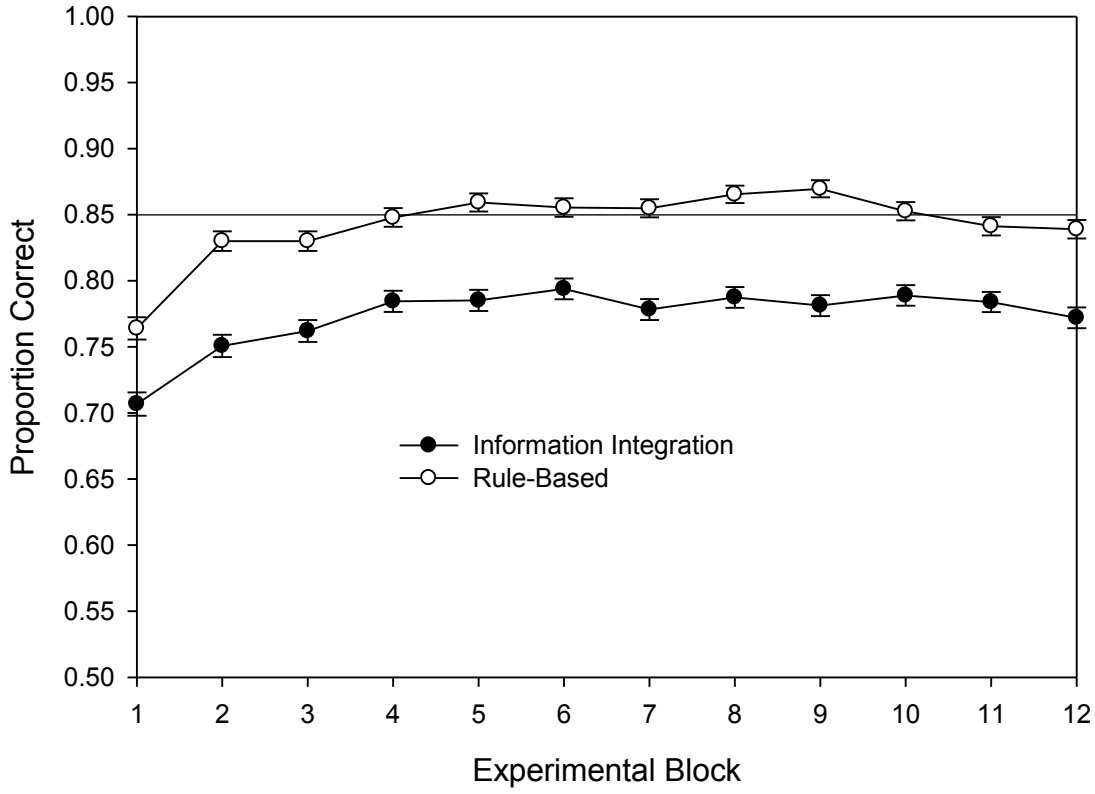

Figure 12. Categorization response accuracy for Experiment 3. The error bars represent the standard error of the mean.

6.2.2 Decision Response Time.

Unlike previous experiments, decision response time was only affected by the Experimental Block, $F(11,638) = 21.46$, $MSE = 1148416$, $p < .001$, $\eta^2_p = .27$. Both Categorization Rule, $F(1,58) = .01$, $p = .91$, and its interaction with Experimental Block, $F(11,638) = 1.42$, $p = .21$, failed to reach statistical significance. These results are presented in Figure 13. The absence of an interaction is readily explained by comparing the results obtained within the present study to those observed in previous experiments. In Experiments 1 and 2, the confidence condition was the only factor that consistently interacted with experimental block whereas all participants provided confidence reports in Experiment 3. When confidence was required, participants mean response times were considerably longer than when no confidence was required which could have led to an absence of difference between these conditions in the current experiment. An interesting result is the apparent discontinuity in the response time figure for primary decision accuracy in the rule-based condition after transfer. The removal of feedback during transfer might have affected the manner in which was confidence processed during the primary decision.

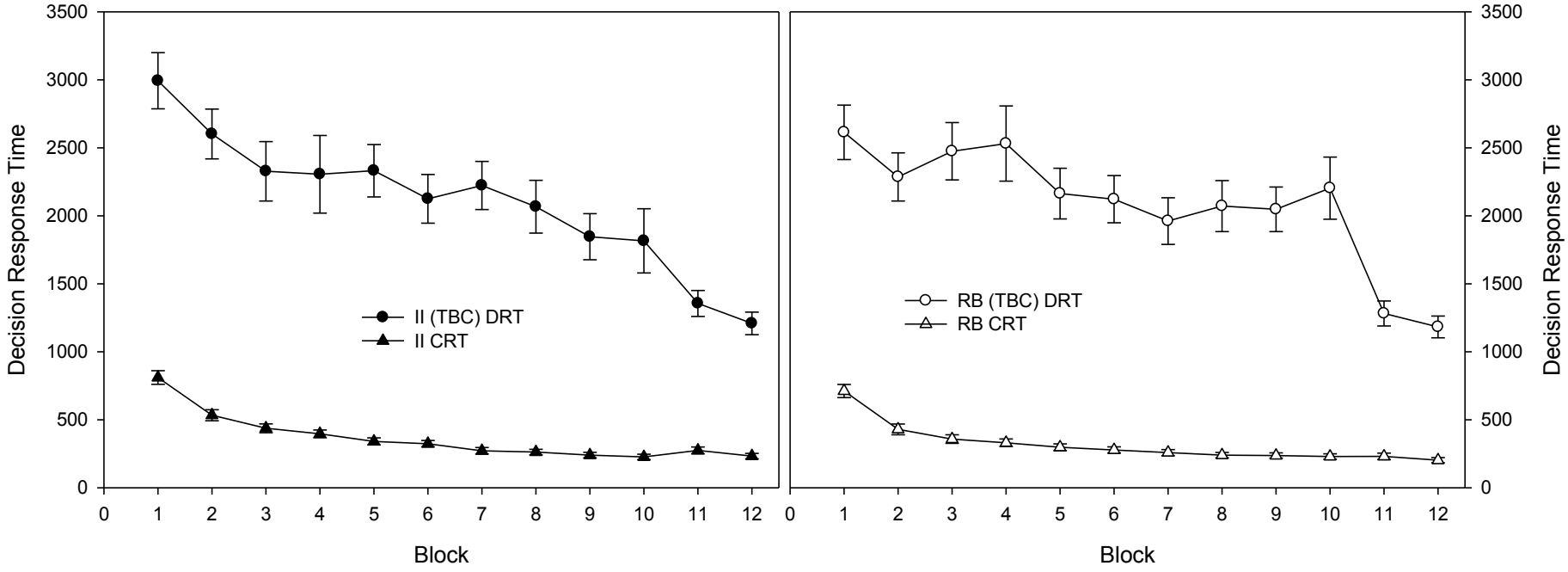

Figure 13. Categorization decision response time (DRT) for trial-level confidence reports and confidence response time (CRT) over experimental blocks. The error bars represent the standard error of the mean.

6.2.3 Confidence Calibration Indices.

6.2.3.1 *Mean Trial-Level*. An analysis of mean confidence reports yielded a significant increase in response confidence across Experimental Block, $F(11,638) = 20.37$, $p < .001$, $\eta^2_p = .26$. A significant effect was also observed for Categorization Rule (see Table 2), $F(1,58) = 4.58$, $p = .04$, but not its interaction with Experimental Block, $F(1,638) = 1.06$, $p = .38$. These results are presented in Figure 14. In general, participants in the information-integration condition reported greater confidence in their responses relative to participants in the rule-based condition. This suggests that the feedback delay not only had the desired effect of reducing performance in the information-integration condition, but it also affected participants' awareness of their performance. A consideration of accuracy and confidence is required to identify whether this reflects an accurate assessment of performance.

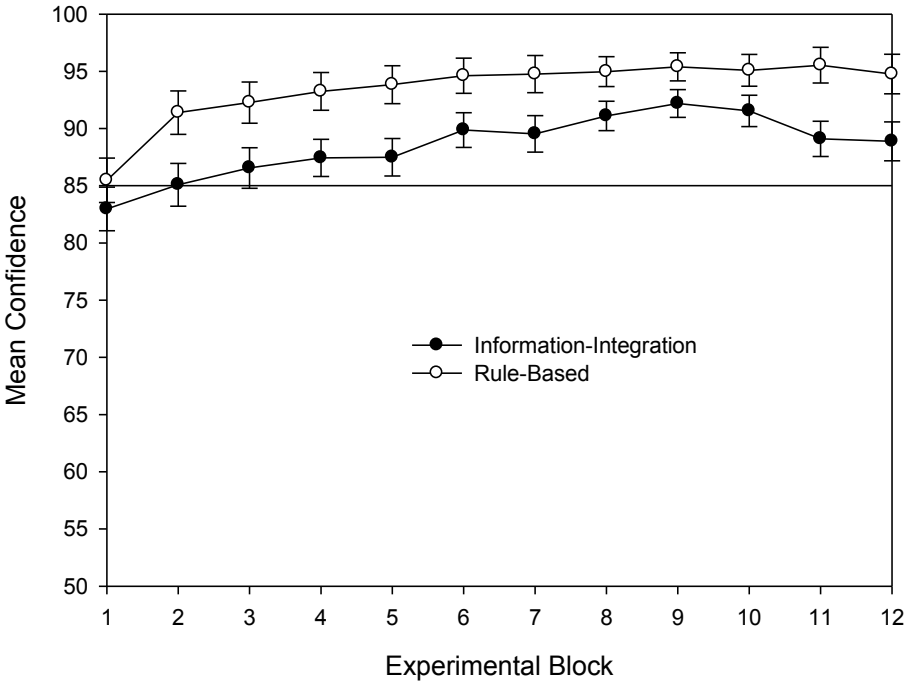

Figure 14. Mean trial-level confidence across category rule conditions. A reference line indicates the 85% performance asymptote. The error bars represent the standard error of the mean.

6.2.3.2 *Trial-Level Subjective Confidence Calibration.* The results of Experiment 3 also provide more evidence for a multiple categorization systems account. They replicated the results of Experiments 1 and 2 as uniform miscalibration was observed for both information-integration or rule-based category structures across Experimental Phases, $F(3, 174) = 4.75$, $MSE = .001$, $p = .016$, $\eta^2_p = .08$. Unlike previous experiments, however, there was a general trend toward increasing calibration after the initial phase of training. A marginally significant difference in performance was also obtained between Categorization Rule (see Table 2), $F(1,58) = 3.29$, $MSE = .001$, $p = .075$, $\eta^2_p = .05$, but no significant effect was obtained for its interaction with Experimental Phase, $F(3, 174) = 1.27$, $p = .28$. An examination of Table 3, however, suggests better calibration in the intermediate and late training phases. Thus, it seems plausible that the resulting change in calibration could simply be due to noise in the data or some as yet unidentified relationship between the requirements of automaticity of stimulus-response mapping early in training and a reduction in performance when feedback is removed during the transfer phase.

6.2.3.3 *Trial-Level Overconfidence Bias*. As in the previous experiments, a significant difference in overconfidence bias was observed between Experimental Phases, $F(3,174) = 4.69$, $MSE = .008$, $p = .006$, $\eta^2_p = .08$, with no significant differences evidenced for Categorization Rule, $F(1,58) = .14$, $p = .71$. In general, a non-significant trend toward greater overconfidence was observed as participants progressed through Experimental Phases (see Figure 11), $F(3,174) = 2.04$, $p = .123$. Qualifying the calibration results, participants exhibited the greatest overconfidence in the transfer phase of the experiment relative to the learning phase. This

suggests that delays in trial-feedback resulted in participants predicting greater performance in the transfer phase than they achieved.

6.2.4 *Trial-Level Confidence Response Time*.

Conforming to the other confidence analyses, the only significant result in the present analyses was Experimental Block, $F(1,638) = 128.57$, $MSE = 35039$, $p < .001$, $\eta^2_p = .69$. Again, no differences were obtained for Categorization Rule, $F(1,638) = 2.20$, $p = .14$, or its interaction with Experimental Block, $F(1,638) = 1.69$, $p = .16$. A straightforward interpretation of the results presented in Figure 13 is that confidence processing occurred in an equivalent fashion in both the rule-based and information-integration conditions.

6.3 Discussion

Taken together, the results of Experiment 3 provide evidence that feedback differentially affects categorization learning systems. First, replicating the findings of previous research (Ashby et al., 2002), participants assigned to the feedback delay condition while learning information-integration category structures did not acquire the category structure as accurately as those participants in either Experiments 1 or 2. These results support the claims that a procedural-learning system that uses response feedback is affected by eliminating or delaying trial feedback whereas a hypothesis-testing is not affected in a similar manner (Maddox et al., 2005). This latter assertion is of particular importance given that the overconfidence bias did not differ between rule-based and information-integration category structures. I will consider the possible implications of this finding further.

As with the previous experiments, these results could be obtained if confidence reports represent a constant bias independent of the primary decision, a constant proportional bias independent of category representations, or the conjoint influences of an explicit representation and response feedback. The observation of reductions in the level of miscalibration in comparison to Experiments 1 and 2 could have resulted from a confidence response bias wherein participants report a constant level of confidence (e.g. 90%) that is independent of any representation. If so, the failure to obtain significant differences in calibration results between rule-based and information-integration category structures could be taken as evidence of such a bias. With a reduction in accuracy in the information-integration condition, an independent bias account would predict an increase in the amount of observed overconfidence bias observed due to a constant level of confidence reported by participants. The similarities in the training phase between Experiment 3 and Experiments 1 and 2 provide evidence against a strong version of such an account.

Experiment 3 also appears to provide evidence against a proportional bias account of this systematic deviation. The performance asymptote, obtained accuracy, and observed level of miscalibration were identical to those obtained in Experiments 1 and 2. A proportional bias account could provide an explanation of these results given that, despite the reduced accuracy in the information-integration condition due to delayed feedback, similar levels of miscalibration across category structures are observed during training. This account, however, has difficulty in explaining the increased overconfidence for participants assigned to the rule-based condition during the transfer phase. If there is a constant proportional bias, miscalibration in the transfer phase should be equivalent given that participants have the same performance asymptote. It is important to recall that during the transfer phase trial-level feedback is no longer provided but

participants maintained the same accuracy in their categorization responses. Thus, the changes in confidence ratings must be related to how the change in feedback affects the feeling of subjective certainty.

A constant bias of confidence resulting from an explicit representation and a proportional bias caused from evidence rescaling provide inadequate accounts of the experimental results. In contrast, adopting a representational account as the basis for confidence reports provides an adequate explanation. The accessibility of the explicit category structure is the primary determinant of response confidence and feedback is used to attenuate this representation. In the rule-based condition, the accessibility and reliance on this representation should be constant across training once the hypothesis-testing system has identified an accurate representation of the category structure. In information-integration conditions, participants should be less inclined to rely on this representation given the proportion of negative feedback that they receive. The representation in the procedural-learning system will be less accessible due to its multidimensional structure and will not produce high levels of confidence. A similar set of processes is likely to be involved in most studies of subjective confidence that use perceptual discrimination and identification tasks, such that participants would need to rescale primary decision evidence from a cognitively impenetrable perceptual system. Once negative feedback is removed, participants might come to rely more on an explicit representation of the category structure as negative feedback no longer attenuates the accessibility of this representation. This account is also supported by calibration results.

Although some support is provided for a representational dissociation account of the representations available to the respective learning system, I sought to obtain additional converging evidence to support these findings. A possible reason for the moderate effects

obtained for dissociable learning systems could be that the accessibility of an accurate explicit representation might produce near-asymptotic feelings of certainty. Participants across all three experiments had representations that accurately reflected the category structure 85% of the time by the end of training. This was further ensured by an elimination of low-accuracy participants using a 65% learning criterion and validated by paired comparisons of obtained performance with the 85% performance asymptote. Under these conditions, it is critical to note that even in the information-integration condition wherein a procedural-learning system should rely on feedback, negative feedback is received on the minority of trials (i.e., ignoring perceptual or decisional noise, this would account for 15% of all trials in the absence of learning the exceptions). The increase in miscalibration in the transfer phase of the experiment suggests an important role for negative feedback on confidence reports.

Ell and Ashby (2006) have suggested that participants *select* which system to rely on during categorization. They found evidence for this idea by fitting one- and two-dimensional models of categorization decision-making to groups of participants that were presented with category structures that varied in their overlap. Replicating early findings (e.g., Ashby et al., 1998), low- or medium-high levels of category overlap resulted in participants initially relying on the hypothesis-testing system and ultimately relying on the procedural-learning system at the end of training. These conditions are effectively replicated in Experiments 1 to 3. At levels of high overlap, the hypothesis-testing system continues to dominate. In their account, the procedural-learning system's failure to sufficiently resolve an accurate representation of the category structure results in a reliance on the representation contained within the hypothesis-testing system. Such conditions would aid in testing the possibility that confidence reports are a product of an explicit representation of stimulus information contained within a hypothesis-

testing system and response feedback. By manipulating the degree of category overlap, the performance asymptote can be uniformly reduced for both rule-based and information-integration category structures while increasing the total negative feedback received by participants. In the rule-based condition, I would still see considerable overconfidence due to an accessible, explicit representation that cannot accommodate exceptional exemplars or process feedback. In the information-integration condition, I would also expect to see greater overconfidence than previous experiments while also seeing a greater differences between rule-based and information-integration conditions due to the reduced diagnosticity of the explicit representation when learning an information-integration category structure. Experiment 4 investigates this possibility.

7.0 Experiment 4

The results of the previous experiments provide evidence for a representational dissociation. If participants have access to explicit representations of the category structure as well as an explicit awareness of the proportion of negative feedback they have received, then differences in subjective calibration, especially overconfidence bias, should be related to the category structure. In Experiment 1, a trend toward greater miscalibration was observed in the rule-based condition leading to overconfidence. In Experiment 3, a trend toward greater miscalibration was again observed, in this case without an observation of significant overconfidence. Taken together, this provides evidence for different representations within procedural-learning and hypothesis-testing systems. It is clear, however, that the conditions leave open explanations other than process dissociation.

A possibility raised in the discussion of Experiment 3 is that participants' confidence reports are primarily determined by the accessibility of an explicit representation of the primary decision. Nevertheless, the relatively high level of accuracy might lead participants to express high confidence in both conditions. In Experiment 1, when participants in the information-integration condition were presented with negative block-level feedback, overconfidence decreased sharply. Although this was not evidenced Experiment 2, it could suggest that participants' awareness of global features of a task can also affect their assessments of their performance. In this case, a highly accessible representation and large proportion of positive feedback would both suggest reliable category structures. Thus, an increase in the total quantity of negative feedback should differentially affect confidence reports in these conditions. Although participants had access to an explicit representation within both information-integration and rule-based conditions, the explicit representation was not diagnostic of the category structure. Again, the introduction of a large amount of category overlap reduced the performance asymptote to 65%, thereby making any simple, verbalizable rule unreliable. If participants had relied on such a rule within the rule-based condition, a general trend toward greater overconfidence should have been evidenced in the rule-based condition due to the reduction in the performance asymptote in both conditions. In contrast, less overconfidence should have been observed in the information-integration condition due to a decreased reliance on the explicit representation. To test these claims, a category structure was adopted that allowed for increased category overlap relative to Experiments 1, 2, and 3. Under these conditions, Ell and Ashby (2006) have demonstrated that categorization accuracy was uniformly reduced in both the rule-based and information-integration conditions while both hypothesis-testing and procedural-learning systems contribute to response selection.

The success of Experiment 4 in dissociating categorization systems will be evidenced by changes in overconfidence bias compared to other experiments. Overall, if confidence is determined by an explicit representation, greater overconfidence should be observed across all conditions. If feedback processing affects the procedural-learning system whereas it leaves the hypothesis-testing system unaffected, changes in overconfidence should be observed within the information-integration condition. Specifically, the participants should express *less* overconfidence bias due to limited accessibility of the implicit representation retained in the procedural-learning system relative to that stored within the hypothesis-testing system.

7.1 Method

7.1.1 Participants and Design.

Eighty-eight participants were recruited from Carleton University and were awarded course credit for their participation in a single 2-hour session. Participants were required to learn a rule-based category structure (defined by frequency; n = 42) or an information integration category structure (defined by frequency and line orientation; n = 46). Half of the participants were assigned to the block-confidence condition (n = 42) whereas the other half were assigned to the trial-and-block confidence condition (n = 46).

7.1.2 Stimuli and Procedure. Stimuli were identical to those used in Experiment 2. The design of the study was also identical to that of Experiment 2 except that the performance asymptote was set at 65%. Unlike Experiment 3, I again used trial- confidence and trial-and-block confidence conditions. Feedback was again presented as in Experiments 1 and 2.

7.2 Results

Unlike previous experiments, preprocessing of data had to proceed in a different fashion given the reduction of the performance asymptote to 65%. Chance (p = 0.5) was selected as an appropriate criterion to ensure that participants were not simply guessing when categorizing stimuli. I found that all participants' classification accuracy was above chance (*M = .660, SD = .071*) suggesting that participants had learned the category structures, $t(102) = 22.86$, $p < .001$. Thus, no participants were excluded in either the rule-based or information-integration conditions. As in previous experiments, there were no significant differences for the learning criterion in comparison to the final learning block (M = .666, SD = .091), $t(94) = 1.71$, p = .09, whereas there was a difference for the last to learning blocks (M = .669, SD = .088), $t(94) = 2.10$, p = .038.

7.2.1 Proportion Correct.

The experimental design was identical to that of Experiments 1 and 2. A 2 (Categorization Rule: rule-based vs. information-integration) x 2 (Confidence Condition: block-level confidence vs. trial-and-block confidence) x 12 (Experimental Blocks: 1-12) mixed-design ANOVA was performed on proportion correct, collapsing across response key. Replicating previous studies using similar category structures (Ell & Ashby, 2006), no interaction was observed between Categorization Rule and Experimental Block, $F(11,924) = .92$, p = .504. Replicating the results of previous experiments (see Figure 15), I observed a significant effect of Experimental Block, $F(11,924) = 10.21$, MSE = .01, $p < .001$, η^2_p = .11, as well as the main effect of Categorization Rule, $F(1,84) = 4.15$, MSE = .06, p = .045, η^2_p = .05. Neither the main

effect of confidence condition, $F(1, 84) = .099$, $p = .754$, nor its interactions with Experimental Block or Categorization Rule were significant (all Fs <.91, all ps > .50).

As in the previous experiments, participants' accuracy increased over blocks of learning trials. In this experiment, I observed that learning was slight better in the information-integration condition relative to the rule-based condition (see Table 2). However, the mean performance (M = .66) across both conditions was nearly identical to the desired performance asymptote (i.e., a learning criterion of .65).

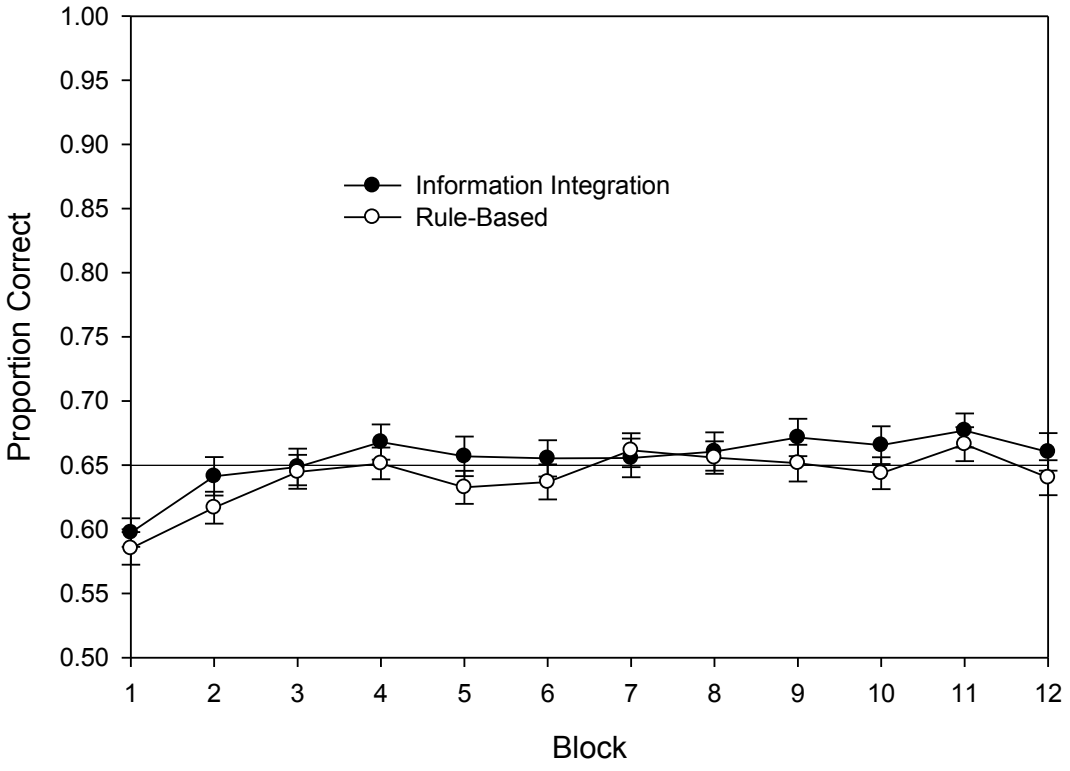

Figure 15. Response accuracy with 65% performance asymptote. The error bars represent the standard error of the mean.

7.2.2 Decision Response Time.

The design of this analysis was equivalent to the one used for accuracy. Supporting the results of the accuracy analysis, a significant effect of Experimental Block was also observed, $F(11,924) = 16.17$, $MSE = 494654$, $p < .001$, $\eta^2_p = .16$. As participants received more blocks of training, they responded increasingly faster. Qualifying this effect, an interaction was also observed between Experimental Block and Confidence Condition, $F(11,924) = 2.73$, $MSE = 494654$, $p = .02$, $\eta^2_p = .03$. A decomposition of this effect revealed a significant simple effect for the block-level confidence condition, $F(11,946) = 4.50$, $MSE = .01$, $p < .001$, as well as for the trial-and-block level condition, $F(11,946) = 14.99$, $MSE = .01$, $p < .001$. A marginal significant effect of category structure was found, $F(1,84) = 2.84$, $p = .10$. The effect of Confidence Condition, $F(1,84) = 1.68$, $p = .20$, and the interaction of Categorization Rule and Confidence Condition, $F(1,84) = 1.85$, $p = .18$ failed to reach significance, however. As in previous experiments, participants were slower to categorize stimuli when they were also required to report subjective confidence (See Figure 16). This pattern can again be interpreted as increasing automaticity across experimental blocks with the requirement of confidence reports resulting in the activation of concurrent process during the primary decision.

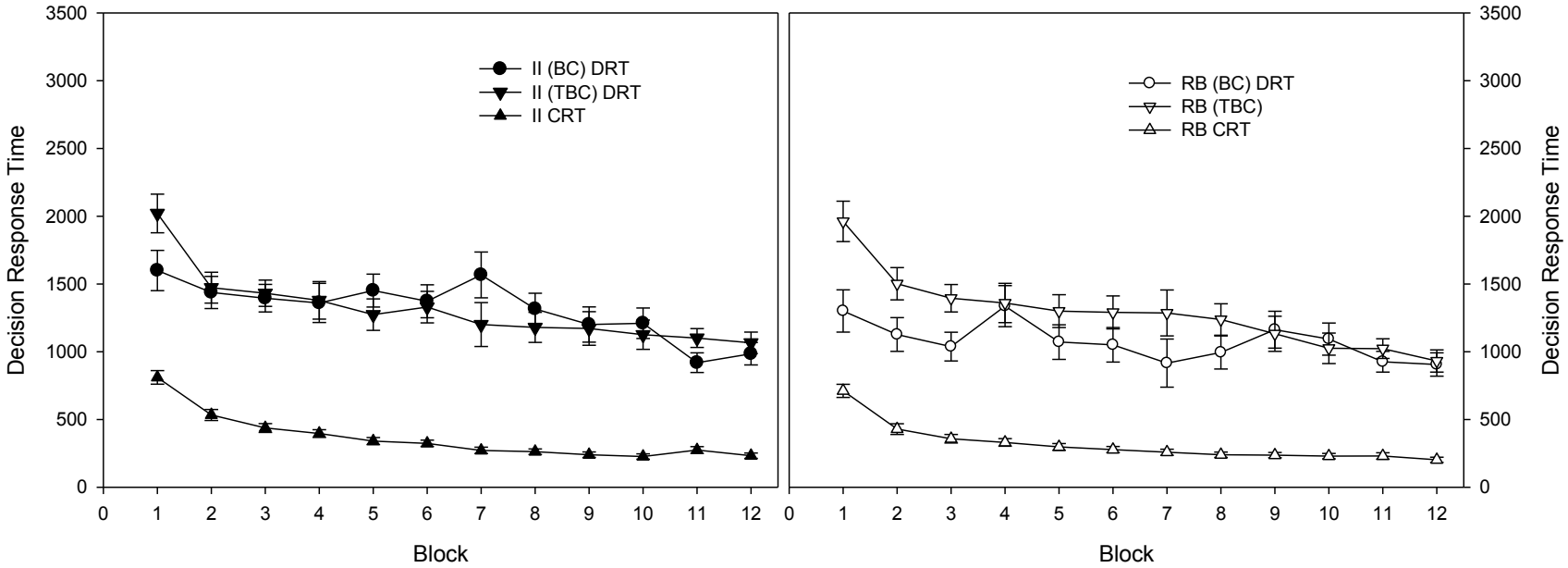

Figure 16. Categorization decision response time for trial- and block-level confidence (TBC) and block-level only confidence (BC). The error bars represent the standard error of the mean.

7.2.3 Confidence Calibration Indices.

7.2.3.1 Mean Trial-Level Confidence. An analysis of mean subjective confidence only revealed a main effect of Experimental Block, $F(11,484) = 4.55$, $MSE = 95.14$, $p < .001$, $\eta^2_p = .09$. The analysis did not reveal a main effect of Categorization Rule (see Table 3), $F(1,44) = 1.44$, $p = .24$, nor its interaction with Experimental Block, $F(11,484) = 1.22$, $p = .30$. Figure 17 demonstrates this trend. Participants expressed the same mean level of certainty across learning conditions. Such a result would be expected if participants had an overall awareness of an increase in their performance over experimental blocks.

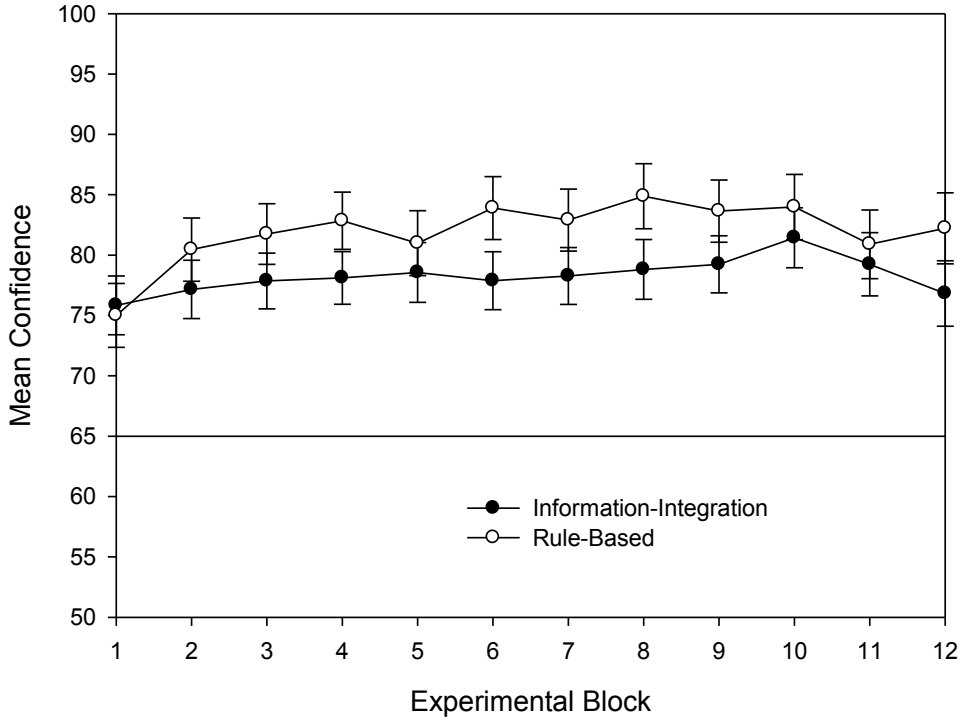

Figure 17. Mean trial-level confidence across category rule conditions. A reference line indicates the 85% performance asymptote. The error bars represent the standard error of the mean.

7.2.3.2 Trial-Level Subjective Confidence Calibration. A 2 (Categorization Rule: rule-based vs. information-integration) x 4 (Experimental Phase: 1-4) repeated-measures ANOVA was conducted. An analysis of subjective calibration did not reveal any difference between Experimental Phase (see Table 3), $F(3,132) = 2.088$, $p = .12$, and obtained only a marginal effect of Categorization Rule (see Table 2), $F(1,44) = 2.83$, $MSE = .00$, $p = .10$. The interaction of Experimental Phase and Categorization Rule was also not found to be significant, $F(3,132) = .428$, $p = .696$. Such a finding is expected given that both accuracy and mean confidence reports reached an asymptote early in training. It suggests that, on a trial-to-trial basis, participants had a similar level of awareness of their performance although there is a trend toward improved calibration from the early phases of training to later experimental phases.

7.2.3.3 Trial-Level Overconfidence Bias. Unlike the calibration analysis, a significant difference in overconfidence bias was observed in the analysis of Experimental Phase, $F(3,126) = 5.36$, $MSE = .01$, $p = .004$, $\eta^2_p = .11$. A significant effect of Categorization Rule was also observed, $F(1,42) = 10.88$, $MSE = .02$, $p = .002$, $\eta^2_p = .21$. These findings suggest that participants' general awareness of the category structure differed between rule-based and information-integration category structures (see Table 2) as well as over experimental blocks of trials. As is evidenced in the nearly additive pattern in Figure 18, participants exhibited greater overconfidence in the rule-based condition relative to the information-integration condition. This finding would be expected if the accessibility of the representation led participants to overestimate their performance in the rule-based condition whereas this effect was attenuated when participants incorporated negative response feedback into their subjective reports. The pattern evidenced in Figure 18 also suggests

that participants' overconfidence bias generally decreased over experimental blocks. This pattern suggests that the rapid generation of a representation within the hypothesis-testing system might lead participants to believe that their performance was in fact better than it was in early phases of the experiment. With additional trials, participants' accuracy eventually caught up to their subjective confidence.

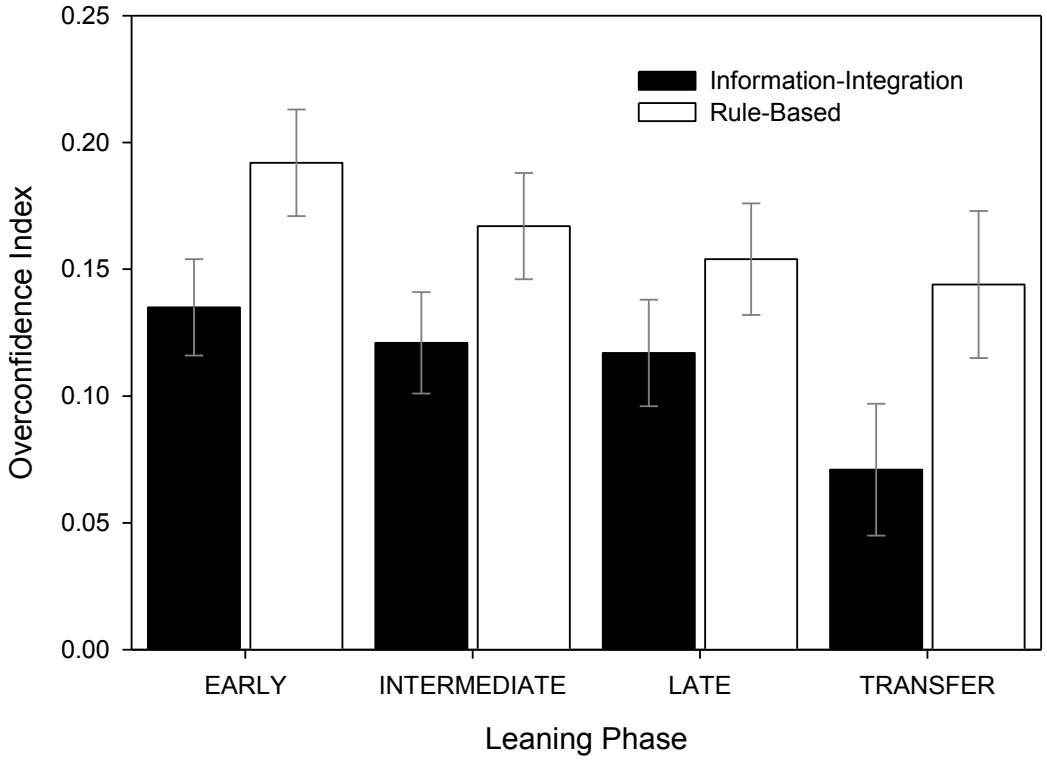

Figure 18. Overconfidence bias for Experiments 4. Maximum overconfidence located at .35 (not plotted). The error bars represent the standard error of the mean.

7.2.4 Trial-Level Confidence Response Time. Confidence response times were analyzed in a repeated-measures ANOVA including Category Structure (rule-based vs. information-

integration) and Experimental Block (1 to 12). Confidence response time was significantly affected by both Experimental Block, $F(11, 506) = 46.34$, $MSE = 43283$, $p < .001$, $\eta^2_p = .50$, and Categorization Rule, $F(1, 46) = 4.56$, $MSE = 157398$, $p = .04$, $\eta^2_p = .09$, but not their interaction, $F(11,506) = .892$, $p = .46$. These results are contained in Figure 16. Conforming to the results of previous experiments, the reduction in confidence response latencies observed here reflects a pattern consistent with increased automaticity during response selection. With more training, participants required less time to assess their certainty in their responses. Similarly, as in the previous experiments, participants in the rule-based condition were faster to report their confidence responses than participants in the information-integration condition. This suggests that a greater amount of additional process was required following categorization for participants in the information-integration condition compared to those in the rule-based condition.

7.3 Discussion

Replicating the findings of previous research (e.g., Ell & Ashby, 2006), the results of Experiment 4 demonstrated that a reduction in performance asymptote had an equivalent effect on performance in both rule-based and information-integration category structures. In rule-based and information-integration conditions, participants learned the category structure at equivalent rates and reached the 65% performance asymptote around the same stage of training (i.e., Block 3). Systematic deviation in response confidence between participants learning rule-based and information-integration category structures would then suggest differences in the representations stored within each category learning system. An analysis of deviations observed within confidence reports qualifies these properties.

The results of Experiment 4 appear to provide further evidence that links both participants' explicit representation of the classification rule and the response feedback that they receive to their confidence judgments. Nevertheless, other less theoretically meaningful phenomena could also account for the results. First, participants in Experiments 1 and 2 could have displayed a response bias where confidence is assigned a constant value greater than of the primary decision (e.g., confidence is 10% greater than the primary decision accuracy). Fortunately, this concern can be dismissed given that as each group reached the performance asymptote. It follows that these results instead suggests that the difference in overconfidence bias that was observed in later experimental blocks could not have been the result of differences in categorization accuracy. This means that differences in subjective confidence ratings must be the sources of the variations observed here. Consequentially, confidence reports must reflect properties of each category structure. If I assume that the processes that gave rise to confidence reports uses information that is accessible to participants, then either an explicit representation of the category structure, response feedback, or both should act as the basis for their responses.

Another related possible explanation of the confidence results could be that participants produced a constant confidence level (e.g., 90% confidence) independently of classification accuracy. Evidence against such an explanation was provided in Experiment 1 because the overconfidence bias differed for the rule-based and information-integration conditions. The results of Experiment 4 also provide further evidence that confidence reports are not simply determined by a constant. If participants simply referred to a constant, participants should express identical levels of confidence (e.g., 90%) in both rule-based and information-integration condition. Instead, the results of Experiment 4 show an increase in response confidence across

experimental blocks. This suggests that while participants are learning the category structure, their confidence is increasing.

An examination of overconfidence bias provides further evidence for a process dissociation account of category learning systems. Although mean confidence and mean accuracy could be independent, overconfidence suggests that there is a systematic relationship. As in the previous experiments, participants in both the rule-based and information-integration condition displayed overconfidence. The relative difference in overconfidence bias between these two conditions is suggestive of differential contributions of the explicit and implicit learning systems. Namely, participants displayed uniformly *greater* overconfidence within the rule-based condition and *lesser* overconfidence over the course of the experiment. If confidence reports are based primarily on information within the hypothesis-testing system, a reduced overconfidence bias in the information-integration suggests that either the explicit representation available within the hypothesis-testing system was perceived as less diagnostic of the category resulting in participants having less confidence in the accessible category structure or that increased negative feedback caused participants in the information-integration condition to encoded more exceptional exemplars into their representation of the category structure. Combinations of these possibilities also present themselves.

The encoding of exceptional exemplars has been taken as the basis for at least one dual-process categorization model (e.g., RULEX, Nosofsky et al., 1994) and the small increase in categorization response accuracy observed in the information-integration condition (shown in Figure 13) might at first be seen as evidence supporting such an account. The obtained differences do not appear to support this explanation given that the gains in accuracy are

considerably less than those observed within subjective confidence. The larger difference evidenced in overconfidence bias suggests that participant did not acquire exceptional exemplars. This finding also conforms to studies that demonstrate that participants generally use linearly separable category structures (e.g., Blair & Homa, 2001). It therefore seems that participants experienced a reduced confidence in an explicit representation or that they were drawing from the implicit representation when assessing their performance. I will consider a combination of these processes.

By definition, participants should not be able to access an implicit representation stored within the procedural-learning system. This suggests that either dual-process accounts that propose to discrete systems are in error (a possibility that I will explore in the Conclusion) or that response feedback indirectly suggests the presence of exceptional exemplars that are not explicitly encoded. On this account, an explicit representation generated by the hypothesis-testing system provides participants with a basis for both categorization and confidence. As training progresses, participants in the rule-based condition find that the use of such a representation is rewarded with positive feedback. On a trial-to-trial basis, they expect that the rule will be accurate but have no other basis for responses. In the information-integration condition, participant attempt to use the explicit representation as the basis for their judgments. Negative feedback that results from their reliance on it indicates that it is in error. In the absence of having a diagnostic category rule that is accessible, participants must adjust their confidence reports on the basis of the recalled level of negative feedback from previous trials. Thus, unlike previous experiments requiring confidence in perceptual discrimination tasks, the non-independence of trial-by-trial decision plays an important role in confidence judgments.

Moreover, if negative feedback in conjunction with an explicit rule causes participants to encode exceptional exemplars, these exemplars might be misclassified on any given trial due to their proximity in perceptual space to other exemplars. As a result, participant would be aware that an exceptional exemplar was located within a given region (i.e., close to category boundary), but might not be able to distinguish it precisely.

In sum, the greater overconfidence evidenced in the rule-based condition relative to the information-integration condition suggests two different categorization systems. Aside from the additional processing time and nominal response bias that arise from rescaling primary decision evidence for a confidence report, systematic differences were evidenced between categorization conditions. These deviations in overconfidence suggest that participants express greater certainty in the rule-based condition. As in other experiments, this can be understood as participants maintaining an explicit representation that they believe is diagnostic of the category structure such as an optimal decision boundary. This boundary fails to account for distributional properties of the category structure leading to overconfidence when applied on a trial-by-trial basis. In contrast, participants in the information-integration condition are less likely to rely on an explicit representation. Participants either reduce a confidence criterion used to judge the amount of accumulated evidence or they attempt to take into consideration exceptional exemplars but cannot use them to categorize stimuli.

8.0 General Discussion

The objective of this thesis was to examine the relationship between categorization learning systems and participants' subjective awareness of their performance. Multiple process accounts of categorization such as COVIS (Ashby et al., 1998) assume that participants have a hypothesis-testing system and a procedural-learning system that compete during response selection. In order to acquire representations of category structures, central processing resources and feedback support learning within the hypothesis-testing system and procedural-learning system, respectively. I assumed that explicit awareness of the properties of category representations would alter subjective confidence over the course of training as well as depend on the nature of the category structure that was presented to participants. In contrast to dual-process models of categorization, many models of confidence processing have remained agnostic as to the representation of the information used to determine certainty (e.g., Baranski & Petrusic, 1998; Ferrel & McGooey, 1980; Vickers & Packer, 1982). Others have assumed that the accessibility of cues provides the basis for judgments of certainty (e.g., Busey et al., 2000; Koriat, 1997). I developed and tested prediction in four experiments derived from both accounts of categorization and confidence processing to investigate whether representational accessibility within the explicit system is the main determinant of confidence reports.

The goal of Experiment 1 was to establish the validity of a paradigm that required participants to learn the category membership of stimuli presented using the randomization technique while also providing trial-by-trial confidence judgments. The selected rule-based and information-integration categories had sufficient overlap to produce a performance asymptote at 85% correct. This asymptote allowed for overconfidence throughout the course of learning (i.e.,

if participants were 90% or 100% confident in a block of trials, they would be miscalibrated). Confidence was also requested after each block of trials. In general, I expected to observe greater overconfidence in the rule-based condition. I hypothesized that participants assigned to that condition should not be able to incorporate the negative feedback or exception exemplars into their representation of category structures if learning was dependent on the hypothesis-testing system. The results of Experiment 1 confirmed this prediction. In both the rule-based and information-integration condition, mean participant accuracy equalled the performance asymptote by the end of training. Similarly, patterns of response latencies in both conditions produced monotonically decreasing functions across learning. These results supports the idea that there is a shift between a system that initially produces algorithmic responses to another system that dominates when a sufficient number of memory traces have been accumulated for automatic retrieval of stimulus-response mappings (e.g., Logan, 1988; see also, Fitts & Posner, 1967). Whereas the differences in certainty might be understood as a consequence of difficulty in learning category structure rather than being attributable to dissociable learning systems (e.g., Newell, Dunn, & Kalish, 2010; Nosofsky & Johansen, 2000), the use of a performance asymptote ensured that participants in both the rule-based and information-integration results conditions attained similar levels of response accuracy. Differences in response certainty then must be a result of *subjective* difficulty which would be a consequence of representation accessibility during response selection. Following from this, difference in confidence reports and calibration indices can instead be interpreted as products of dissociable information processing systems.

An examination of confidence calibration indices lent support to the hypothesis that there are dissociable categorization learning systems. Following the first phase of training, participants in the rule-based condition reported greater overconfidence relative to those in the information-integration condition. These results indicated that participants attained optimal performance, but that they overestimated their success in categorizing stimuli. This finding would be expected if participants failed to account for the category overlap resulting from exception exemplars. Under such conditions, the accessibility of the representation of the hypothesis-testing system could be contrasted against the comparatively inaccessible representations of the exception exemplars. Such a finding also replicates the finding that that participants form an optimal classification boundary in order to categorize stimuli (e.g., Ashby & Gott, 1988; Blair & Homa, 2001).

Experiment 2 removed block-level feedback to ensure that the results of Experiment 1 were not a product of a global adjustment of criterion used for performance monitoring rather than a result of different representations used within the hypothesis-testing and procedural learning systems. The results of this experiment generally replicated those of Experiment 1. Overconfidence was observed in both rule-based and information-integration conditions. Unlike Experiment 2, however, the greater overconfidence observed in the rule-based condition during the transfer phase was not significantly different than that observed in the information-integration condition. This could be attributed to the removal of block-level feedback thereby suggesting the global criterion change that I sought to avoid. More plausibly, given the comparative level of accuracy for participants in both the rule-based and the information-integration conditions relative to those in Experiment 1, it is as likely that participants simply learned the category structure faster in the information-integration condition. Given that nearly

identical methods were used in Experiments 1 and 2, the differences in results appear to be a result of individual differences in learners.

Experiments 3 and 4 sought to increase overconfidence in order to replicate the findings of Experiment 1 as well as introduce greater disparity between accuracy and confidence to demonstrate representational dissociation. In Experiment 3 delayed feedback reduced learning within the procedural-learning system. If the explicit representation used to inform confidence reports is unaffected by reductions in performance, a greater level of mean overconfidence should be obtained within the information-integration condition. Participants in the rule-based condition reached the asymptote whereas participants assigned to the information-integration condition did not because feedback was delayed. These findings replicate the results of other experiments that have used delayed feedback as a means to interfere with the procedural-learning system during information-integration learning (e.g., Maddox et al., 2003). The additional observation of overconfidence in these results suggests that the accessibility of the representation within the hypothesis-testing system leads participants to have greater confidence in their responses. The constant level of overconfidence observed in Experiment 3 and Experiment 1 additionally suggests that the overestimation of performance is proportional to the obtained performance.

The first three experiments obtained a pattern of overconfidence consistent with my hypothesis. Specifically, the greater level of overconfidence bias under conditions in which the hypothesis-testing system suggests that the accessibility of category representation during response selection is the primary determinant of subjective awareness. Thus, the dominance of the implicit, procedural-learning system during category response selection could explain the

reduction in overconfidence bias given the reduced accessibility of this representation. Experiment 4 sought to increase overconfidence by increasing the number of exceptional exemplars presented to participants. If an explicit representation cannot accommodate exceptional exemplars and the hypothesis-testing system is comparatively unaffected by negative feedback, increasing the number of exceptional exemplars and negative feedback should increase the magnitude of the overconfidence bias. Using Medium-High category overlap parameters used by Ell and Ashby (2006), I reduced mean performance to 65% correct for both the rule-based and information-integration conditions. Replicating previous results, participants learned the rule-based and information-integration category structures equally well. Moreover, the greatest level of overconfidence was obtained in this experiment suggesting that neither exceptional exemplars nor negative feedback was incorporated into the explicit representation that guided categorization performance. Taken along with the results of the three previous experiments, these findings conform to a dual-process account of categorization learning systems (e.g., COVIS, Ashby et al., 1998).

The results of the present experiments lead to several conclusions concerning the nature of the category representation retained within the two hypothesized learning systems, the relationship between these learning systems and confidence, and the utility of the confidence methodology that was employed to examine subjective awareness of performance. I will consider each of these below and attempt to reconcile several candidate models of category representation and categorization learning systems with the literature on subjective awareness.

8.1 Multiple Representations and Confidence Processing

Although the primary purpose of the present thesis was not to address the respective assumptions of models of confidence processing, the necessity of interpreting the calibration indices requires a clear understanding of how confidence reports are produced. In so doing, a greater understanding of the representation of response certainty will suggest a relationship with the representation of the category structures. If representations of the category structure used to categorize stimuli and that used to determine subjective confidence differ, then different category structures should produce different levels of overconfidence bias. As is illustrated in Figure 19, when representations used to categorize stimuli and report confidence are compatible because they are both rule-based, then subjective confidence should be high. Under conditions in which the representations used to categorize stimuli and report confidence are not compatible (i.e., when classification requires the procedural-learning system), subjective confidence should be low. Thus, prior to discussing the findings of my analyses of confidence responses, I will first consider how my results relate to extant models of confidence processing. Moreover, some of the results obtained within these experiments provide crucial insight into the nature of representations and processes involved in reporting confidence.

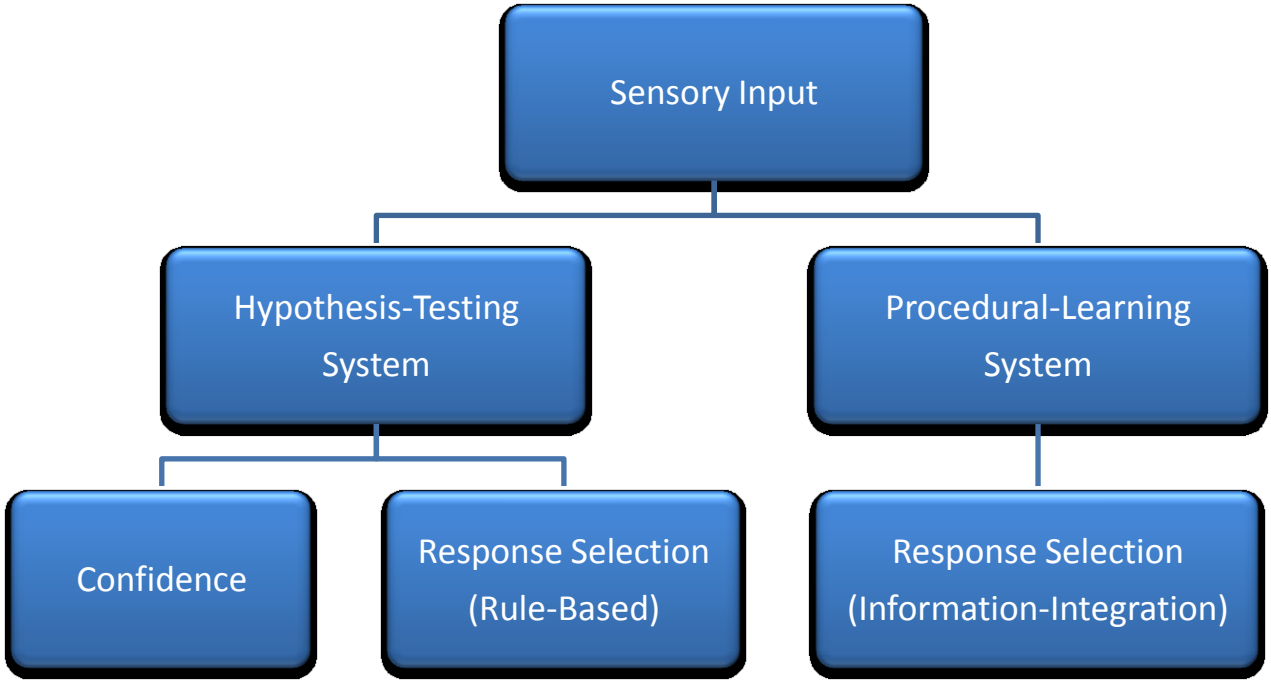

Figure 19. Stimulus processing during categorization and subjective confidence. Sensory information is activated and propagates within the system activating representations of exemplars within the hypothesis-testing system and procedural-learning system. Both categorization systems activate a motor response during response selection with the system that reaches a criterion level of activation responding. In addition, an explicit system produces a confidence response that can incorporate global feedback.

8.1.1 Direct-Scaling Models of Confidence. Direct-scaling accounts of confidence processing represent a broad class of models that suggest that only a single cognitive process is required to complete the primary decision response selection and compute confidence. These models provide a straightforward account of the subjective experience of certainty wherein the strength of a representation (i.e., the magnitude of evidence favouring the dominant response) is directly

scaled into a confidence report. Ferrel and McGooey (1980) proposed the first detailed account of a direct-scaling model based on signal-detection theory. The primary decision was the results of using the strength of evidence relative to a decision criterion to select a response. Thus, as the distance between an exemplar and the category boundary increases, the strength of evidence supporting its membership in a given category also increases. In such an account, subjective confidence is determined by obtaining the distance from the decision criterion without any additional computation required (see also Balakrishnan & Ratcliff, 1996; Norman & Wickelgren, 1969; Treisman & Faulkner, 1984). Thus, direct-scaling models must predict that the distance from the category boundary is the determinant of a participant's confidence in the present experiments. Greater distance from the category boundary must yield greater subjective confidence. It is interesting to note that similar claims have also been made in the metamemory literature (e.g., Donaldson, 1996; Hart, 1967; Wixted & Stretch, 2004).

Recent categorization studies have used and interpreted confidence reports in a manner that is consistent with direct-scaling models (e.g., Estes, 2004; Rehder, 2003; Ziori & Dienes, 2008). A useful point of comparison, however, is a study conducted by Balakrishnan and Ratcliff (1996). It assessed a direct-scaling model of subjective confidence in the context of categorization task that employed the randomization technique. In that study, two participants performed two unidimensional categorization tasks using line-lengths and numeric magnitude. In each task, stimuli were randomly selected from a unidimensional stimulus distribution. Following each classification, participants provided confidence reports on a 10-point scale from 1 (certainty in Category A) to 10 (certainty in Category B). Each participant received a total of 13 hours of training (including a discrimination task in addition to the two categorization tasks).

Using a signal detection analysis of the hit rate, their results demonstrated a high degree of correspondence between obtained performance and mean confidence response. Moreover, their participants were reasonably well calibrated showing trends that are more consistent with a pattern of underconfidence. Using a strength-based model of confidence reports, they were able to demonstrate that their results could be accounted for by a single-process model of response selection and confidence. Again, according to the single-process model of confidence provided in COVIS, this degree of correspondence would be expected in the output of an implicit, procedural-learning system was used to scale confidence. Namely, Ashby et al. (1998) can be taken as claiming that the amount of activation (i.e., strength of response) within the procedural-learning system is used to determine confidence in later stages of training.

Although I did not attempt to fit Balakrishnan and Ratcliff's (1996) response model to my confidence responses, my results do not appear to be accommodated by such a single-process model. In contrast to Balakrishnan and Ratcliff's (1996) results, my unidimentional rule-based category structure produced considerable overconfidence, a pattern of performance that neither matches Balakrishnan and Ratcliff's (1996) results nor one that is predicted by the direct-scaling model of confidence referenced in the context of COVIS. A single-process model could conceivably be capable of modeling these affects with the addition of noise to the response selection process or attention to irrelevant features. In either case, these additional sources of information would bias confidence responses. Such an account would also conform to the suggestion that the observed patterns of performance in rule-based and information-integration conditions are due to difficulty (Newell et al., 2010).

If the introduction of a biasing source of information could account for the differences in confidence responses observed in the present study, however, an unanswered question would be left open: Why would more noise be present in the rule-based condition in comparison to the information-integration condition? A differential increase in noise or division of attention suggests that an additional process mediates confidence response selection. Alternatively, even if information decay, or leakage, occurs between the primary decision and confidence report (Usher & McClelland, 2001), a plausible explanation must be offered as to why there would be more information decay in one condition relative to another. Moreover, after participants in both rule-based and information-integration condition have reached the performance asymptote, the observed differences in confidence reports and primary decision accuracy can no longer be a product of noise within the categorization process. Taken together, the results of the present study strongly suggest that additional model parameters would be necessary for a single-process account of confidence to model the findings of the present study.

8.1.2 Rescaling Models of Confidence and the Process of Rescaling.

The observed differences in overconfidence bias obtained in the rule-based and information-integration conditions suggest that confidence processing is associated with a set of operations other than those used in categorization response selection. As noted in the previous section, single-process models of confidence judgment such as those based in SDT appear to be inadequate in accounting for the pattern of overconfidence observed in the present study. Similarly, increases in categorization response time with the requirement of confidence suggest a set of operations other than those used to categorize stimuli is required. Thus, the representation

obtained from the category learning systems must somehow be altered in order to report confidence through a process of rescaling accumulated evidence into a confidence report. Understanding the translation of evidence from the primary decision to confidence report provides insights into the connection between the representation available to both the hypothesis-testing system and that used to report confidence.

Issues associated with confidence scaling have emerged comparatively recently, likely resulting from initial assumptions that levels of confidence merely reflected additional decision criterion, but not processing (Balakrishnan & Ratcliff, 1996; Ferrel & McGooey, 1980; Pleskac & Buseymeyer, 2010). In contrast, rescaling models of confidence make the minimal assumption that confidence reports require rescaling, or re-representation, of primary decision information. For instance, Audley (1960) described confidence as an inverse function of the amount of vacillations between "runs" of evidence accumulated for response alternatives (e.g., ABBB is associated with greater confidence than ABABBB) until a criterion number of successive accumulation events is reached. Alternatively, Vickers and Packer (1982) and Baranski (1991; Baranski & Petrusic, 1998) suggested that subjective confidence is determined by computing the difference between accumulated evidence for the response alternatives (dominant and non-dominant responses) or the quantity of nondiagnostic evidence, respectively. Thus, the output of a categorization process must be used as the input to a confidence process in order to select an appropriate confidence response. Such an additional process should result in increases in response time. Prior to automatization, response selection is typically characterized as an effortful process (e.g., McCann & Johnston, 1992; Szmalec & Vandierendonck, 2007; Szmalec, Vandierendonck, & Kemps, 2005). If response selection occurs at a central processing stage,

requiring similar resources for primary decision and confidence response selection, then the requirement of confidence reports would delay the completion of the primary task. An observation of an increase in categorization response time when confidence is required would suggest related effortful response selection operations are required. I observed such evidence within my study.

First, across four experiments, I found longer responses latencies when confidence was required relative to when participants simply categorized stimuli. This replicates findings in the sensory discrimination literature that have been used to argue for the requirement of additional rescaling operations (e.g., Baranski & Petrusic, 2001). Baranski and Petrusic (1998) have used such findings to claim that this represents a concurrent processing of the primary decision and confidence reports. Moreover, Baranski and Petrusic (2001) have also demonstrated that a decrease in response times for confidence reports across blocks, which suggested the development of automaticity. Indeed, if confidence reports relied solely on a post-decisional locus of processing that drew from information accumulated over the course of the primary decision, it would be difficult to see why this would increase *categorization* decision response time. It would also difficult to understand why participants would intentionally engage in response selection of confidence reports during the primary decision. Hence, these response times indicate that confidence reports are sampling the same information as the categorization operations, but that they acting independently. It could even be speculated that confidence judgments require the generation of a representation other than that used for the primary decision response selection.

Second, differences in categorization response accuracy and confidence could be attributed to perceptual or decisional biases. For instance, as I noted above, it might be argued that the information-integration categorization condition was inherently more difficult (Newell et al., 2010), and that this fact gave rise to reduced overconfidence relative to the rule-based categorization condition. Such an account is in line with the Hard-Easy Effect (e.g., Lichtenstein & Fischhoff, 1977). It identifies a robust phenomenon that easy task typically yield underconfidence whereas hard ones typically yield overconfidence (see also Baranski & Petrusic, 1994). Participants in both rule-based and information-integration conditions reached the performance asymptote and continued to exhibit differences between response confidence and decisional accuracy in Experiments 1, 2, and 4. Thus, even if differences in acquiring category structure existed early in training, objective differences in difficulty were no longer present after the performance asymptote was reached. If objective difficulty was equivalent following the performance asymptote, perceived differences in subjective difficulty appear to be the source of uncertainty experienced by participants. Subjective difficulty could result either from rescaling information available from the primary decision or could instead be a result of difficulty in accessing information used to categorize stimuli. As I have shown above, there is an associated difficulty with rescaling that result from scale parameters (e.g., Schoenherr & Petrusic, 2011). Given that the same confidence scales was used in both conditions, and given that both conditions had equivalent accuracy, an alternative source of subjective difficulty is required.

Taken together, the changes in response time as well as differences in overconfidence observed across experimental conditions once accuracy has been equated suggest that additional rescaling processes are required. If the process of rescaling that results from the requirement of

confidence reports varies in difficulty, this suggests that the representational content of the primary decision information must differ. Difficulty, in these terms, would then be understood as the amount of representational change that is required to alter a representation used in the primary decision. For instance, in Figure 18, sensory input activates both categorization systems. If participants' responses are already dominated by the hypothesis-testing system, little representational change would be required. If instead the procedural-learning system dominates response selection, the nondominant representation contained within the hypothesis-testing system must be rescaled in order to produce a confidence response. Thus, there should be greater difficulty associated with providing a confidence report when the representation used to respond is the nondominant representation. This focus on representational content leads to a final consideration: the source of the information used during confidence processing.

8.1.3 Source of Information. Instead of assuming that the general difficulty of a task can result in changes in confidence (e.g., Lichtenstein & Fischhoff, 1977), the alternative explanation proposed here is that difficulty is a result of differential levels of representational accessibility used to complete the categorization task in each condition. This conforms to the account provided by Dawes (1980) with recent evidence also provided by Kvidera and Koustaal (2008). In Dawes's (1980) study, participants were found to exhibit greater overconfidence in their responses to general knowledge questions relative to their performance in a perceptual discrimination task. Dawes (1980) attributed this difference to the greater accessibility of general knowledge compared to the relative inaccessible of encapsulated perceptual processing systems. In his account, the ability to retrieve general knowledge coupled with the cognitive

impenetrability of sensory processes leads participants to overconfidence and underconfidence, respectively. Recently, Kvidera and Koustaal (2008) also obtained similar results when they used the same stimuli in a general knowledge and a discrimination task. Participants were presented with names of countries that varied in terms of their in geographic area. The names were presented in either a large or small font. Participants were then asked to judge which of the two stimuli was smaller or larger in terms of either geographic area or font size corresponding to a general knowledge task and a perceptual judgement. Kvidera and Koustaal (2008) observed uniform overconfidence in both general knowledge task and the perceptual judgement task but less overconfidence in the perceptual judgements of font size. This finding provides a pattern that is similar to that obtained by Dawes (1980). Moreover, these results suggest that different information processing systems will produce levels of confidence that correspond to representational accessibility. Moreover, as categorization requires memory systems, I will now turn to evidence from dual-process of memory that can inform our understanding of categorization learning systems.

The majority of evidence supporting the independence between information used to respond to stimuli and that used to report confidence comes from the metamemory literature (e.g., Busey et al. 2000; Koriat & Ma'ayan, 2006; Roediger, Rajaram, & Geraci, 2007). Again, initial strength-based accounts (e.g., Hart, 1967) have given way to cue-based accounts (e.g., Koriat, 1997; Koriat & Ma'ayan, 2006). Koriat (1997), for instance, proposed that participants retrieve items from memory and use the accessibility of this information to report their certainty. Later studies by Koriat and Ma'ayan (2006) went on to show that either encoding or retrieval

cues determine response certainty depending on whether those reports are provided shortly after encoding or after delayed in retrieval, respectively.

The "remember/know" paradigm in studies of recognition memory has also demonstrated a dissociation between confidence reports related to "remembering" and "knowing" (e.g., Roediger, Rajaram, & Geraci, 2007). These tasks require participants to indicate that they "remember" test stimuli when they have explicit knowledge whereas they should indicate "know" when the test stimuli only appear familiar. Strength-based accounts of metacognition in the remember/know paradigm have assumed that stimulus strength is the sole determinant of confidence resulting in higher confidence in "remember" relative to "know" judgments. Instead, results have demonstrated that confidence reports might be better understood as affected by the nature of the representation that is accessed rather than primary decision retrieval accuracy (e.g., Rajaram et al., 2002; for a review, see Yonelinas, 2002). For instance, Koriat and Ma'ayan (2006) observed that metacognitive assessments of performance differed depending on whether they were required shortly after encoding or following a long delay. They took this as evidence that certainty was judged on the basis of either encoding or retrieval cues, respectively. This suggests that multiple sources of representational information might need to be concatenated in order to arrive at an assessment of certainty.

In the context of confidence models, source information concatenation can simply be understood as rescaling. Initially encapsulated perceptual processes must first resolve visual information. Participants might experience some awareness of perceptual information, but it is not likely to become an integrated, explicit representation (Dehaene & Changeux, 2011). In the case of confidence reports for information-integration category structures, implicit information

requires an additional association with response scale categories relative to when an individual is presented with a rule-based category structures. Eventually, the automatization of the rescaling process reduces response latencies. The differences in response times in rule-based and information-integration condition appear to support such a process: the function describing RTs when confidence is required in the rule-based condition generally reaches an asymptote faster than that in the information-integration condition indicating that automaticity occurs earlier in learning for the rule-based condition. This would occur if confidence reports increase primary decision response time as a result of concurrent processing (e.g., Baranski & Petrusic, 1998) and became progressively automatic with the continued reporting of confidence (e.g., Petrusic & Baranski, 2001). This process, however, is not addressed by all models of confidence processing.

8.2 Confidence Processing in a Multidimensional Categorization Task.

Confidence processing models typically focus on a single, direct-scaling process (Ferrel & McGooey, 1980) with more recent models simply requiring that the evidence accumulation process continue after primary decision response selection (e.g., Pleskac & Busey, 2010). Other models have suggested a secondary set of operations (e.g., Audley, 1960; Vickers & Packer, 1982), but they failed to account for an alterable processing locus, a feature indicative of a separable response selection process. One model of confidence processing could be adapted to explain the present findings. As noted above, Baranski (1991; Baranski & Petrusic, 1998) suggested that a doubt-scaling process could give rise to confidence reports.

Using the basic proposal of the doubt-scaling model, information that is not diagnostic of the primary decision is stored in a separate accumulator representing uncertainty. Thereafter, it is

scaled into a confidence report. In the context of the present study, what constitutes nondiagnostic information would change depending on the diagnosticity of the explicit representation. In the rule-based condition, a participant might initially assume that she is guessing. The explicit representations that are generated and tested by the hypothesis-testing system would quickly produce an optimal classification rule. In the absence of perceptual information to the contrary, two perfectly linearly separable categories defined by such a rule would be perfectly diagnostic of category membership. In these conditions, confidence reports based on this explicit representation would result in perfect calibration. If it is further assumed that the hypothesis-testing system exhibits little feedback dependency, negative feedback should not affect confidence reports but would affect categorization accuracy. With the provision of category overlap, accuracy would be reduced resulting in overconfidence. This is in fact the pattern of observed results. Specifically in Experiments 1, 2, and 3 on 15% of trials, participants would receive negative feedback. Thus, if confidence is based on an explicit representation that is not affected by negative feedback, I would anticipate an equivalent amount of overconfidence (i.e., 0.15). Moreover, in Experiment 4 in which participants received negative feedback on 35% of trials, I observed a similarly greater level of overconfidence. The obtained values for these three experiments reflect such a bias that appears to reflect the proportion of negative feedback received by participants. These findings conflict with a direct-scaling, strength-based account of confidence processing adopted by Ashby et al. (1998) in the context of COVIS.

In the information-integration condition, I again expected overconfidence due to reliance on an explicit representation. In this condition, I assume that participants would find that their explicit representation is less diagnostic of category membership and would show less

overconfidence bias as a result. Over time, participants should become increasingly aware of the inaccuracy of the explicit rule resulting in a general decrease in overconfidence overtime. Such a finding is also compatible with the underconfidence-with-practice effect observed in the metamemory literature (e.g., Koriat, et al. 2002). Rather than obtaining underconfidence, I instead generally observed a reduction in overconfidence bias. Again, these results were obtained although a statistically significant effect was only evidenced with high levels of category overlap in Experiment 4. In this experiment, less overconfidence was observed in the information-integration condition than in the rule-based condition and overconfidence decreased over time. These results suggest that confidence reports are primarily determined by an explicit representation of the category structure and that they are adjusted with the provision of feedback. It is important to note that such an account requires additional representational assumptions than that provided by Baranski and Petrusic (1998; Baranski, 1990), but such a modified doubt-scaling account would appear to be an appropriate model. For instance, participants could store exemplars as nondiagnostic information leading to increased uncertainty. Alternatively, evidence accumulators could retain the number of feedback events rather than perceptual information or combine abstract representations of both perceptual and episodic information.

8.2.1 Shared Neurological Basis of Categorization and Performance Monitoring.

The difference in overconfidence observed within the rule-based and information-integration conditions suggests that confidence report have utility in dissociation categorization learning systems. The foregoing discussion observed that an increase in response latencies with the requirement of confidence as well as the greater response times observed within the

information-integration condition relative to the rule-based condition suggested that the representation used to report confidence is related to that contained with the hypothesis-testing system. Further evidence for this proposed relationship is also evidenced in neurological correlates between performance monitoring and processes involved in hypothesis-testing system. In the context of a categorization task for which participants are presented with feedback after making a decision, a reasonable assumption is that confidence processing shares similarities with error monitoring. Error monitoring consists in the activation of a response system that can identify the commission of an incorrect response (e.g., Rabbit, 1966). Confidence reports appear to require monitoring the current state of accumulated evidence prior to receiving feedback, as evidenced in increased in DRT relative to when no confidence reports are required. If confidence processing and error monitoring constitute similar operations, then the activation of neuroanatomical regions associated with error monitoring should also be active in situation where participants are required to report confidence.

In an error monitoring paradigm, participants are provided with a task in which they must provide a rapid response after which they can indicate whether they have committed an error (e.g., Rabbit, 1966). The extent to which participants were correct in detecting errors and whether responses following errors have longer response latencies are assessed to infer properties of an individual's monitoring ability. Thus, a participant's task in this paradigm is to assign a correct (100% confidence in correct) or incorrect (100% confidence in error) label to their previous response. Schoenherr and Petrusic (2011) compared scales similar to those used in error monitoring (0, 100% confidence) and confidence reports (0% through 100% confidence) in order to determine the equivalence between confidence response scales. His results suggested

that an error monitoring scale and the confidence response scale produce similar levels of calibration. This suggests that a similar rescaling process occurs when either type of reports is required. Identifying this similarity at a behavioural level allows for a consideration of the neuroanatomical regions associated with error monitoring and control (for reviews, see Botvinick et al., 2001; Taylor et al., 2007).

Of primary importance for the present set of experiments, error monitoring and rule-based category learning appear to activate similar neuroanatomical regions (Ashby et al., 1998; Ashby & Valentin, 2005). Studies of the localization of function of the respective learning systems in COVIS (Ashby & Valentin, 2005; Ashby et al., 2007; Nomura et al., 2007) have found evidence for a role of the anterior cingulated cortex (ACC) and dorsal lateral prefrontal cortex (dlPFC). With sufficient activation of the ACC, the dlPFC is involved in selection of the appropriate rule and maintains activity within an anterior loop (e.g., medial dorsal nucleus). Whereas the findings of these studies conform to earlier conceptualization of COVIS provided by Ashby et al. (1998), when considered alongside studies of error monitoring, COVIS's account of confidence processing and its relationship to the learning systems becomes inconsistent. Again, COVIS itself assumes bottom-up projections for an implicit learning system in order to generate confidence reports. In contrast to COVIS's predictions, studies of error monitoring have found that activation within the ACC is correlated with error monitoring whereas activation with the dlPFC is correlated with control cognitive functions (Taylor et al., 2007).

Given the behavioural similarity of error monitoring and confidence processing as well as the similar patterns of activation of error monitoring and the hypothesis-testing system, this suggests an alternative relationship between confidence and categorization learning system than

that provided by Ashby et al. (1998). Specifically, if confidence reports are influenced by similar processes as error monitoring, then the ACC and dlPFC should also be involved when monitoring performance and reporting confidence. As noted by Taylor et al. (2007), error detection implies a level of subjective awareness with some studies showing greater activation of error detection regions (in this case, pMFC) when participants reported awareness of an error (also see, Hester et al. 2005). Similarly, Kerns et al. (2004) demonstrated that activation of the dACC predicted activity of the dlPFC on subsequent trials. This finding was interpreted as an indication for greater involvement of executive control on trials after an error has been detected. For my purposes here, the interaction of the ACC and dlPFC in both performance monitoring tasks and rule-based categorization tasks appears to be more than coincidental. If processing occurs within the same or neighbouring regions, then it seems reasonable to infer that there would be less difficulty (if any) associated with the kind of representation change implied by rescaling of confidence reports. Moreover, I would further suggest that comparatively shorter response latencies in the rule-based condition relative to the information-integration condition likely reflects reduced activity of the ACC to alter the representation within the dlPFC due to a reduced need for monitoring changes in rule-based category structure. This is also supported by findings that response latencies increase on trials following errors (e.g., Rabbit, 1966).

It is important to note, however, that the need to select between rules associated with stimulus dimensions and ACC activation alone is not likely to be sufficient to describe response confidence. For instance, it could be assumed that confidence responses might be predicted by the inverse of the level of activation in the ACC. One area of research that considers the role of generating response alternatives is the underdetermined response task such. It requires

participants generate a verb when presented with a noun (Andreason, et al., 1995; Petersen et al. 1988). In tasks such as these that require the effortful generation of a response, the ACC is active. When the responses are well-learned or specified in advance, however, the ACC shows less activation. In a categorization task, the ACC might be more active due to additional monitoring resulting from category overlap. However, if the ACC alone were responsible for subjective confidence, I would predict less activation in later phases of training resulting in increased confidence which would result in overconfidence given the performance asymptote. Although there were some increases in overconfidence in Experiments 1-3, there was a decrease in overconfidence across conditions in Experiment 4. This suggests that ACC activation alone cannot account for confidence responses. Possible avenues for further investigation might be to examine the activation of other areas associated with the hypothesis-testing system (e.g., medial dorsal nucleus and association cortex) that are believed to maintain a representation of the category structure (e.g., Ashby et al., 1998).

8.3 Models of Categorization: Evidence from Confidence Reports and Process Dissociation

The analysis of confidence reports provides a new source of evidence for dual-process accounts of categorization. In general, the patterns of miscalibration observed here indicate that the representation used to report subjective confidence and to report categorize exemplars were informed by different sources of information. Specifically, I observed higher levels of overconfidence for those participants learning a rule-based category structure relative to those who learned the information-integration category structure. Greater overconfidence suggests that participants' explicitly learned category structures did not also represent stimulus variability. It is

possible that during the process of rescaling primary decision accumulated evidence could have decayed (Usher & McClelland, 2001). If so, as I noted above, it is less clear how information decay could have resulted in the increases in confidence that led to the overconfidence in only the rule-based condition observed in the data. Such a pattern of differential decay in these conditions would still suggest that some form of representational change is required. More plausibly, it would seem to be the case that different stimulus representations of the category structure were available to two categorization systems.

The evidence suggests that there are two categorization learning systems that differ in terms of their accessibility. In brief, the hypothesis-testing system's accessible representations appear to exert major influence on confidence reports with information that influences the procedural-learning system, such as negative feedback, exerting comparatively less influence. In general, this conforms to the predictions of COVIS concerning the categorization systems, although COVIS does not provide a sufficient explanation concerning their relationship with confidence processing. This shortcoming prompts a consideration of alternative dual-process categorization models.

8.3.1 RULe-plus-EXception (RULEX) Model.

RULEX (Nosofsky et al., 1994; Nosofsky & Johansen, 2000) assumes that participants will first attempt to use a simple rule to categorize stimuli, a property shared with COVIS. Unlike COVIS, however, following the adoption of a categorization rule, exceptional exemplars will be encoded as separate representations in memory. The main issue with adapting RULEX to the present study is that it does not provide a basis for response confidence. In the absence of an

explicit account of response confidence, a number of confidence models could be adapted to RULEX including the dissociable relationship proposed above. This issue is illustrated when considering the general difficulty that exemplar-based models have in acquiring linearly separable category boundaries (e.g., Blair & Homa, 2001). Thus, representations of both rule and exemplar could function as determinants of response confidence. Yet, if participants can encode rules and exemplars, it would seem to suggest that subjective calibration should be high, a pattern not observed in the present study. Instead, the results of the present study suggest that participants adopted an optimal classification rule in both rule-based and information-integration conditions and did not encoded exception exemplars. On these grounds, I will conclude that RULEX does not adequately account for my data set. Alternative models that use rule and exemplar-based representations (e.g., ATRIUM; Erickson & Kruschke, 1998) are also likely to have the same issue as well as those models that simple assume exemplar-based learning (e.g., ALCOVE; Kruschke, 1992) without additional parameters.

8.3.2 Supervised and Unsupervised Stratified Adaptive Incremental Network (SUSTAIN).

Although not a dual-process account, SUSTAIN (e.g., Love & Medin, 1998; Love et al., 2004) provides another possible explanation for the relationship between categorization and confidence reports. A basic assumption of SUSTAIN is that clusters of features constitute a category and that there is response competition between clusters with a bias toward simple solutions. Each cluster represents a single dimension. Thus, like COVIS, SUSTAIN first attempts to categorize stimuli based on a unidimensional structure. When a single cluster is insufficient to successfully categorize stimuli, additional clusters are recruited. Unlike COVIS,

however, SUSTAIN does not provide an account of the relationship between explicit and implicit representations. Hence, if SUSTAIN is used to make theoretical predictions about confidence processing, then the focus must be on its computational characteristics. The most obvious possibility is that the number of clusters that compete to determine the model's classification decision is inversely related to the model's subjective confidence (Equation 6 in Love et al., 2004). If this assumption is taken to be true, then certain predictions can be drawn and compared to the findings of the present study.

As Love et al. (2004) note, humans find unidimensional solutions easier than multiple dimension solutions. Even in cases of non-linearly separable category structures used in the present set of experiments, participants are still capable of learning an optimal category structure such as in the present study (e.g., Ashby & Maddox, 1990; Blair & Homa, 2001). In the rule-based condition, SUSTAIN would recruit a single cluster to categorize the stimuli whereas in the information-integration condition, two clusters would be required. As cluster competition would be lower in the rule-based condition, this would result in greater confidence relative to the information-integration condition because the latter requires more clusters to represent the category structure thus resulting in lower confidence. Unlike the account of confidence processing provided in COVIS, the predictions of SUSTAIN are generally supported by the present data set. Interestingly, at first sight, this might suggest a novel, indirect-scaling model of confidence that uses only primary decision information.

One reason to assume that SUSTAIN's account of confidence processing is somewhat incomplete pertains to the nature of exceptional exemplars. Namely, SUSTAIN assumes that clusters are recruited when a single dimensional cluster fails to predict performance. In both rule-

based and information-integration conditions, exceptional exemplars were included in the category structures. On average, given that the participants rarely exceeded the performance asymptote, few if any exceptional exemplars appear to have been encoded. This leads to a straightforward assumption that the competition that results from exceptional exemplars relative to the category clusters is negligible. It is still an open question why participants would express any uncertainty if only a single cluster was used. This might suggest that participants are attempting to recruit clusters, but that they are not sufficiently active to affect categorization performance. Nevertheless, even if this speculation is correct, it would suggest that the confidence process must incorporate another source of information.

What makes SUSTAIN somewhat problematic with regard to the present research question is its failure to explicitly address participants' awareness of the category structures. This is obviously not an issue for COVIS. For instance, Ashby et al. (1998) note that an explicit, hypothesis-testing system dominates early stages of learning and that this falls within the scope of explicit awareness. Over the course of learning, responses within the implicit learning system result in automaticity and eventually dominate response selection. SUSTAIN adopts a similar approach over the time course of learning. Namely, that single clusters should be used initially with multiple cluster recruited as learning progresses. Rather than response competition occurring between learning systems, it occurs between clusters. Thus, the hierarchical interactive relationship posited by COVIS and supported by neuroscientific evidence is absent in SUSTAIN. What is clear from the present study is that both COVIS and SUSTAIN offer plausible learning mechanisms, but account for only certain aspects of the present findings. Further attempts need

to be made to bring together the adaptive flexibility of SUSTAIN with the neurological underpinnings identified by COVIS.

The participants that took part in the present experiments also produced some intriguing, unexpected behaviours that support dissociable learning systems. Following completion of the transfer phase, participants were required to provide a general description of each category as well as use a computer program to generate a "typical" member of each. After the completion of each task, participants provided their confidence report. In the first three experiments for which the performance asymptote was 85%, the experimenter was sometimes asked by the participants to clarify what the rule was that described the relationship between the two categories. In the fourth experiment with the 65% performance asymptote, the majority of participants (admittedly, a rough estimate) asked whether there had been any rule that accurately described the category structure. They frequently stated that it did not seem like there was a means to accurately separate the two categories. As in the other experiments, the experimenter explained that there had been a significant amount of category overlap that prohibited acquiring the category structures perfectly. He then drew their attention to the description sheet they had just returned on which they typically rated their confidence at 80% for the entire experiment! When confronted with this incompatible information, the participants' reactions were typically of moderate shock – they clearly had not realized that they had a clear image in mind that was not associated with their performance on the task nor did their generated prototypes take into consideration the negative feedback they had received. Granted, this evidence is at best anecdotal, but it is quite suggestive.

8.3.3 Alternative Accounts

The preceding discussion has assumed that the differences observed in categorization and confidence responses was a result of different representations contained within the hypothesis-testing system and the procedural-learning system. In contrast to this account, it might be the case that the representation used to determine confidence ratings is not a product of the hypothesis-testing system, but of another explicit process. For instance, Gazzaniga (2011) has suggested that the verbal processing of stimuli in the left hemisphere has given rise to an "interpreter" module. According to Gazzaniga, this module consistently attempts to explain events in the world and generates "stories" to this end (see also, Nisbet & Wilson, 1977). Whereas this module is typically adaptive and might be seen as the source of humans' propensity to seek causal explanations, these narratives need not be correct. Gazzaniga demonstrated this in an experiment with a split-brain patient (for an earlier account, see Gazzaniga, 1967). Two distinct target images (i.e., a chicken foot and a winter scene) were separately presented to his left and right visual fields, respectively. The patient was then asked to point to a thematically related picture for each target. When he saw the chicken's foot, the patient correctly explained that the picture of a chicken was the correct answer. He incorrectly explained, however, that he had selected a shovel in answer to the other target because it was needed to clean the chicken coup. The interpreter module was not aware that the right hemisphere had selected a shovel in response to the winter scene. Hence, it generated a story that coherently linked the chicken, the chicken's foot, and the shovel.

Gazzaniga's account is not necessarily inconsistent with the relationship between the hypothesis-testing system and confidence that I have suggested, and could be used to explain

overconfidence bias. Namely, the interpreter might receive the products of microconscious processes (i.e., Dehaene & Changeux, 2005; Zeki, 2003) and create a coherent representation that appears internally consistent. According to this account, participants might be "replaying" the evidence accumulation process rather than accessing the representation contained within the hypothesis-testing system. Thus, confidence reports might instead be a product of our explicit awareness of the evidence accumulation process rather than sampling from the representation of the information used to categorize stimuli.

Differentiating between the use of an explicit representation or an explicit process as the basis for confidence cannot be directly examined here. It is interesting to note, however, that participants' block-level confidence across all experiments (not reported here) demonstrated little overconfidence bias relative to their trial-level confidence. At a minimum, these results suggest that participants sampled different information to generate confidence at the trial- and block-level. In block-level confidence reports, better calibration could be the result of participants retrospectively assessing the ratio of positive to negative feedback received in the previous block. Alternatively, participants could have taken multiple implicit strength-based representations of stimulus information which the interpreter module translated into an explicit representation of the category structure. In this case, separate verbalizable and nonverbalizable explicit representations would be available to participants via the interpreter module and hypothesis-testing system, respectively. Participants might then use the nonverbalizable representation to categorize stimuli and the verbalizable representation to report confidence. Further study of verbal descriptions and the responses provided in the visual generation task would help adjudicate between these claims by demonstrating the extent to which explicit

knowledge was consistent across these reporting modalities. I will instead focus on confidence reports as my primary means to assess subjective awareness.

8.5 Methods for Assessing Subjective Awareness

The categorization results provided here conform to the predictions made within the COVIS model (Ashby et al. 1998). In contrast to these predictions, I have claimed that my results do not support relationship between categorization learning systems and confidence reports suggested by Ashby et al. (1998). Prior to accepting my claim that the hypothesis-testing systems and confidence processing use the same representation, I will consider two alternative models for examining subjective awareness: Ziori and Dienes's (2006, 2008) zero-correlation and guessing criteria, and Paul et al.'s (2011) uncertainty response methodology.

8.5.1 *Zero-Correlation and Guessing Criteria*. Ziori and Dienes (2006, 2008; See also Dienes& Berry, 1990) used a confidence scale equivalent to the one employed in the present study. They suggested that guessing and zero-correlation criteria could be derived from response accuracy and confidence reports to examine implicit and explicit knowledge, respectively. The guessing criterion was derived by determining whether response accuracy is greater than chance when participants indicate that they are 50%. If their performance is greater than chance despite their subjective report, this suggests that they maintain implicit knowledge of the category structure. Such a finding would be equivalent to underconfidence using the calibration indices of the present study. The zero-correlation criterion examines whether performance and subjective confidence covary. Namely, if both response accuracy and confidence increase together, this suggests that participants have access to some explicit knowledge. This pattern is equivalent to some degree of calibration using the calibration indices of the present study. There is a crucial

distinction between the measures used by Ziori and Dienes (2006, 2008) and those used in the present study, a point I will describe below.

Despite the apparent face validity of these measures, there are at least two concerns that prohibit a straightforward interpretation of Ziori and Dienes's (2008) results. The first concern is that the zero-correlation measure examines the relationship between mean confidence for correct and incorrect responses within an entire experimental condition. By adopting this approach, it is unclear whether participants are aware on any given trial of the knowledge used to categorize stimuli. In particular, it is conceivable that over a block of trials, participants' performance might increase as well as their confidence but that these might not occur on the same trial. While a correlation expresses the extent to which this co-occurrence is observed, it appears more principled to consider participants' response on a trial-by-trial basis. In contrast to Ziori and Dienes's (2006, 2008) studies of confidence processing have used confidence calibration which examines the difference between a subjective probability rating and a participant's accuracy on a trial-by-trial basis. On these grounds, confidence calibration would appear to be a more principled means of determine whether, on average, individuals are aware of factors influencing their performance.

A second concern follows from the guessing criterion which limits Ziori and Dienes's (2008) examination of the confidence-accuracy relationship to the guessing (50%) category. For instance, Dawes and Mulford (1994) have noted the possibility that regression toward the mean can create artifactual underconfidence within the guessing category and overconfidence in the certainty category. If presented with a two-alternative forced-choice task, participants might be required to rate their confidence on a scale of 50% (guess) to 100% (certainty). When they report

50% confidence, participants' responses are likely to regress to above-average performance (e.g., p(cor) > 0.5) independently of any systematic relationship between the decision-making and confidence process. Observing underconfidence in the 50% category might at first appear to suggest a lack of awareness of performance associated with implicit knowledge. Instead, it could simply be the case that imperfect scaling of responses at the extreme end of the response scale produces underconfidence without reflecting changes in awareness. Thus, satisfying the guessing criterion alone is not likely to be a convincing demonstration of implicit knowledge.

Additional concerns about the reliance on a guessing criterion stem from other studies of confidence reports (e.g., Baranski & Petrusic, 1994; Kvidera & Koustaal, 2008) that have observed underconfidence in the 50% confidence category while at the same time observing a general overconfidence. Moreover, task dependencies have also been observed such that difficult tasks typically produce overconfidence and easy tasks produce underconfidence. This is the *Hard-Easy Effect* (e.g., Lichtenstein et al., 1982). Thus, a more robust measure of the influence of implicit knowledge would be to examine the overall *over/underconfidence bias*, the difference between mean confidence and accuracy across all confidence categories. Given the zero-correlation criterions focus on global performance and neglect of trial-by-trial performance as well as the guessing criterions focus on detecting an underconfidence bias in a single confidence category rather than over all categories, I conclude that Ziori and Dienes' (2006, 2008) measures do not provide as valid a measure of explicit and implicit knowledge in categorization tasks in comparison to the calibration indices used in the present study.

8.5.2 *Uncertainty Response Methodology*. Paul et al.'s (2011) study provides another instructive point of comparison to the present experiments. They employed the randomization

technique as well as assessing uncertainty associated with categorization responses. Using Gabor patches, participants were either presented with rule-based or information-integration category structures. Training proceeded in an identical fashion to that employed in the present study with categorization responses followed by feedback. In the experimental phase, participants were then told that they could indicate whether they were uncertain about a stimulus's category membership by pressing the '?' key. When a short familiarization phase preceded training and experimental blocks in Experiment 1, no differences were observed between categorization rules. When no training in the use of uncertainty responses was provided in Experiments 2 and 3, a greater frequency of uncertainty responses were produced.

The greater uncertainty obtained by Paul et al. (2011) in Experiments 2 and 3 is comparable to the reductions in overconfidence bias observed in the present study. Thus, the increase in uncertainty responses is consistent with the Underconfidence-With-Practice (UWP) effect noted previously. Despite these results, two issues need to be considered. One issue with uncertainty responses is that it forces participants to make a discrete decision concerning their certainty. It is quite possible that participants are biased to under- or overreport their level of uncertainty. For instance, participants might exhibit a regression toward the mean such that their "uncertain" responses exhibited greater than chance performance (e.g., Dawes & Mulford, 1994). This is similar concern to that associated with the guessing criterion used by Ziori and Dienes (2006, 2008). A further issue might arise from embedding the judgment of uncertainty in the primary task rather than deriving it post-decisionally as in the present study. Recall that rather than categorizing a stimulus and reporting confidence, Paul et al. permitted participants to select an alternative response. In this way, reporting uncertainty could simply be understood as

identifying a third category intermediate to the two primary categories under investigation. This might represent a task that requires a different set of processing resources and might not reflect a metacognitive process. To exclude this possibility, a condition would need to be introduced that would allow participants to identify a third region of perceptual space associated with these stimuli (i.e., Category C) in order to ensure equivalency between tasks. In this case, participants would effectively need to learn two parallel decision boundaries in order to partition perceptual space into three categories. In the absence of such a condition, it is not clear how to interpret the uncertainty response methodology used by Paul et al.

The results of the present study appear to support an account of confidence process that implies that uncertainty judgments entail a secondary process that is not necessarily commensurate with the categorization task. Evidence for this also appears to be evidenced in Paul et al.'s (2011) study. Relative to when participants received training in the use of uncertainty responses (Experiment 1), participants in Experiments 2 and 3 performed more poorly in the information-integration condition when uncertainty responses were required. From this, one can infer that their results could have been the product of the need to automate response scaling of the uncertain responses during the categorization task. No response times were provided in their study that would eliminate this possibility. Although associated with concerns over rescaling and the requirements of a secondary set of processes, the confidence report methodology used in the present experiments appears to provide reliable ratings across training and reduce interference due to stimulus-response re-mapping while allowing for an assessment of subjective awareness across multiple categorization structures.

8.6 Representational Assumptions and Further Directions

One of the central concerns of categorization research has been the nature of the representation used to categorize stimuli. Initial assumptions that categorization required verbal rules containing necessary and sufficient conditions have given way to similarity-based accounts using summary representations, exemplars, and category boundaries (for a review, see Goldstone, Kersten, & Carvalho, 2012). Rather than maintaining that there is a single kind of representation retained within a general learning system, researchers have suggested multiple representational systems (e.g., Ashby et al., 1998; Nosofsky et al., 1994) as well as multiple representations (e.g., Ashby & O'Brien, 2005; Minda & Smith, 2001; Blair & Homa, 2001). If the dual-process account examined here is correct, this allows for the elimination of several possible constraints on category representations.

An explicit system is generally restricted to low-dimensional representations (e.g., a single stimulus dimension). Prototype and decision-bound models offer plausible category representations given that they represent either an average of features or a criterion rule that can be used to segment perceptual space. Mathematically, such models can be shown to be equivalent at a computational level (Ashby & Maddox, 1990). The resource limitations of a dynamic hypothesis-testing system would appear to make these representations ideally suited to allow participants to group stimuli together in coherent categories. More complex category structures, however, require people to retain an even greater quantity of information about stimuli. For instance, any natural category contained within a scientific or folkbiological system

has a large number of dimensions (e.g., tail, ears, fur colour, eye shape, size) that would appear to be outside the scope of the limited capacity of a hypothesis-testing system.

The complexity of the stimuli within the external world makes the procedural-learning system far more important for organisms living in an environment defined by environmental regularities. Another difficulty with prototype and decision-bound models is that participants do not appear to retain distributional properties of categories associated with exemplars (e.g., Nosofsky, 1986; Stewart & Chater, 2002). Although an optimal classification rule might inform participants that two groups of stimuli differ, the boundary itself cannot retain properties concerning the density of the respective distributions or their spread. Participants do appear to retain many of these properties leaving an important question: Are complex category structures inaccessible to participants in order for them to make verbalizable reports? The answer must be "no". Moreover, it would appear to be rash to claim that a hypothesis-testing system and a procedural-learning system are entirely dissociable when methodological considerations are taken into account. Participants must be capable of detecting both dimensions in the two-dimensional stimuli used here and, indeed, they often identified these dimensions correctly when requested. Thus, like all tasks, categorization cannot be thought to be process-pure (Jacoby, 1991; Ziori & Dienes, 2006, 2008). Multiple neuroanatomical structures are likely active during both rule-based and information-integration conditions and determine response selection. This suggests that COVIS provides a first approximation of how category learning systems work independently but further refinement of such a model should additionally focus on integrating these two systems.

A straightforward means to reconcile a competition-based model of categorization with neurobiology is to adopt an approach comparable to that used within the working memory literature. Initially, multi-store models that drew sharp distinctions between short- and long-term memory (Atkinson & Shiffrin, 1968) gave way to working memory with a distinctive central executive and passive retention stores (Baddeley & Hitch, 1974). Later models of working memory that included more modal executive function (e.g., Baddeley, 2000) illustrated the distributed nature of processing and lack of a single locus for controlled processing. More recently, researchers have instead adopted accounts based on the level of neurological activation (e.g., Anderson, 1983; Conway & Engle, 1994; Cowan, 1995; for a review, see Miyake & Shah, 1999). On these accounts, information in long-term memory receives increasing levels of activation. Items that are highly active are retained in working memory whereas a single item can be retained in the focus of attention.

Category structures are likely acquired in a manner similar to that outlined in activation-based accounts of working memory. Stimulus information from the association cortex is available to neuroanatomical regions such as the ACC for the generation of rules, but the resulting representations are limited in the number of stimulus dimensions that define them. This available information would constitute representations contained within the rule-based system. Concurrently, humans can retain irrefutably complex category structures at the level of their subjective awareness (e.g., birds, clinical disorders, and engines). An individual's attention can capture more of these features with greater levels of experience and ultimately develop expertise. Expertise can be defined as the ability to bring clusters of features together rapidly into the focus of attention to engage in rapid pattern recognition with these patterns evoking other patterns and

associated behavioural schema. The ubiquity of the explicit system in reasoning about novel tasks and the dependency of the explicit system on the implicit system make them on some level inseparable. For reasons such as these, it might be more appropriate to be agnostic concerning the nature of the representation of category structure (e.g., Rosch, 1973; Rosch & Mervis, 1981) and instead focus on the extent to which people are aware of the category structures that they can use accurately. The confidence report paradigm presented here offers a method for the analyses of these representations. Confidence reports can be used to identify differences between category representations used to respond in a task and those that they are aware of at a trial- and block-level. Using these methods researchers should be capable of reaching insights concerning the nature of representations that participants have access to, the condition in which they are aware of the features that define these representations, as well as how learning systems interact to give rise to behaviour.

References

Ahn, W. (1990). Effects of background knowledge on family resemblance sorting. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 149-156). Hillsdale, NJ: Erlbaum.

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, 16, 81-121.

Andreason, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, L., Boles Ponto, L. L. & Hichwa, R. D. (1995). Remembering the past: two facets of episodic memory explored with positron emission tomography. *American Journal of Psychiatry, 152,* 1576-1585.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, 8, 629–647.

Arbuckle, T. Y.,& Cuddy, L. L. (1969).Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, 81, 126-131.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, 114, 632-656.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, 14, 33-53.

Ashby, F. G., Boynton, G., & Lee, W. W., (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, 55, 11-27.

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., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception & Performance*, 16, 598-612.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 50-71.

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

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30, 666-677.

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

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

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

Ashby, F. G., & Valentin, V. V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 547– 572). New York: Elsevier.

Audley, R. J. (1960). A stochastic model for individual choice behaviour. *Psychological Review*, 67, 1-15.

Baddeley, A. (1966). The capacity for generating information by randomisation. *Quarterly Journal of Experimental Psychology,* 18, 119-129.

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 615-633.

Baranski, J. V. (1991). *Theories of Confidence Calibration and Experiments on the Time to Determine Confidence.* Unpublished doctoral dissertation, Carleton University, Ottawa, ON.

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. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 929-945.

Baranski, J. V., & Petrusic, W. M. (2001). Testing the architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, 55, 195-206.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, 11, 211-227.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General,* 127, 55–68.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology,* 39, 585–609.

Besner, D., & Coltheart, M. (1979).Ideographic and alphabetic processing in skilled reading of English. *Neuropsychologia*, 17, 467-472.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics*, 54, 75–81.

Blair, M., & Homa, D. L. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition, 29,*1153–1164.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., Cohen, J. D., (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652.

Budescu, D. & Wallsten, T. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behaviour and Human Decision Processes*, 46, 240-263.

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.

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

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), Cognition and Categorization, (pp. 169-211). Hillsdale, NJ: Erlbaum.

Bruner, J., Goodnow, J., & Austin, A. (1956). *A Study of Thinking*. New York: Wiley.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA.: MIT Press, Bradford Books.

Chan, C. (1992). *Implicit Cognitive Processes: Theoretical Issues and Applications in Computer Systems Design.* Unpublished doctoral thesis, University of Oxford.

Cleereman, A. Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Science*, 2, 406-416.

Cohen, J. D., & O'Reilly (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planner and prospective memory. In Brandimonte, M., Einstein G. O. & McDaniel M. A., (eds.), *Prospective Memory: Theory and Applications*, 267-296. Mahwah, NJ: Erlbaum.

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology*, 45, 523–538.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17,297–338.

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 189-202.

Dawes, R. M. (1980). Confidence in intellectual vs confidence in perceptual judgments. In E. D. Lanterman & H. Feger (Eds.) *Similarity and Choice: Papers in Honor of Clyde Coombs* (pp. 327-345). Bern Hans Huber.

Dawes, R. M. (1994). *House of Cards*. New York: Free Press.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organizational Behavior and Human Decision Processes*, 65, 201-11.

Dehaene, S., & Changeux, JP. (2005). Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattentional blindness. *PLoS Biology*, 3, 910-927.

Dehaene S, Changeux JP. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70, 200–227.

Dehaene, S., & Changeux, JP., Naccache, Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, 10, 204-211.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, 4, 3-23.

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21, 1322–1338.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, 24, 523-533.

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

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

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, 18, 500-549.

Estes, W. K. (2004). Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. *Psychonomic Bulletin & Review*, 11, 1041-1047.

Evans, J. St. B. T., Barston, J., Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407, 630–633.

Fernandes, M. A., & Moscovitch, M. (2000). Divided attention and memory: Evidence of substantial interference effects at retrieval and encoding. *Journal of Experimental Psychology: General*, 129, 155-176.

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

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33, 238–244.

Fitts, P. M., & Posner, M. I. (1967). *Human Performance*. Belmont, CA: Brooks/Cole.

Frith, C. D., Friston, K. J., Liddle, P. F., & Frackowiak, R. S. J. (1991a). A PET study of word-finding. *Neuropsychologia, 29,* 1137-1148.

Frith, C. D., Friston, K. J., Liddle, P. F., & Frackowiak, R. S. J. (1991b). Willed action and the prefrontal cortex in man: A study with PET. *Proceedings of the Royal Society of London, Series B, 244,* 241-246.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*, 23–30.

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition*, *19*, 617–623.4

Gazzaniga, M. S. (1967). The split brain in man. *Scientific American*, 217, 24-29.

Gazzinga, M. S. (2011). *Who's in Charge? Free Will and the Science of the Brain*. New York: Harper Collin.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506-528.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2012). Concepts and Categorization. In Weiner, I.B, Healey, A.J., & Proctor, R.W. (Eds.) *Handbook of Psychology, Volume 4, Experimental Psychology*, 2nd Edition (pp. 607-630). New York, NY: Wiley.

Goldstone, R. L., Steyvers, M., Spencer-Smith, J. & Kersten, A. (2000). Interactions between perceptual and conceptual learning. In E. Dietrich & A. B. Marman (Eds.), *Cognitive Dynamics: Conceptual Changes in Humans and Machines*, (pp. 191-228). Lawrence Erlbaum and Associates.

Griffin, P., & Tversky, A (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411-435.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, 15, 55-71.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208–216.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356-388.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review*, 16, 824-831.

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

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, 18, 186-201.

Herdt, G. (1994). *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. New York: Zone Books.

Hester, R., Foxe, J. J., Molholm, S. Shapner, M., & Garavan, S. (2005). Neural mechanisms involved in error processing: A comparison of errors made with and without awareness. *NeuroImage*, 27, 602– 608.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411-428.

Homa, D. (1984). On the nature of categories. *Psychology of Learning & Motivation*, 18, 49-94.

Hull, C. L. (1920). Quantitative aspects of the evolution of concepts. *Psychological Monographs*, XXVIII.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98, 352-376.

Huttenlocher, J., Hedges, L. V., & Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. *Psychological Review*, 95, 471-484.

Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513-541.

Jacoby, L. L.,& Brooks, L. R.(1984). Nonanalytic cognition: Memory, perception and concept learning, in *The Psychology of Learning and Motivation*, Vol. 18. ed. Gordon H. Bower, New York: Academic Press, 1-47.

Jacoby, L.L., Toth, J.P., Yonelinas, A.P., & Debner, J.A. (1994). The relationship between conscious and unconscious influences: Independence or redundancy? *Journal of Experimental Psychology: General*, 123, 216-219

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology,* 45, 482–553.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgement: A sampling model of confidence in sensory discrimination. *Psychological Review*, 104, 344-366.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss, Giroux.

Karmiloff-Smith, A. (1992). *Beyond Modularity.* MIT Press/Bradford Books.

Kay, P. (2005). Color categories are not arbitrary. *Cross-Cultural Research*, 39, 39-55.

Keele, S.W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003).The cognitive and neural architecture of sequence representation. *Psychological Review*, 110, 316–339.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217-273.

Kéri, S. (2003). The cognitive neuroscience of category learning. *Brain Research Reviews*, 43, 85-109.

Kerns J. G., Cohen J. D., MacDonald A. W. III, Cho R. Y., Stenger V.A., & Carter C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, 303, 1023-1026.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, 112, 500-526.

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

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General,* 124, 311–333.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General,* 126, 349–370.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* 6,107–118.

Koriat, A. & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language*, 52, 478–492

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experiment Psychology: General*, 131, 147–162.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, 10, 294–340.

Kvidera, S., & Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual decisions. *Journal of Behavioral Decision Making*, 21, 253–281.

Lacroix, G. L., Giguere, G., & Larochelle, S. (2005). The origin of exemplar effects in rule-driven categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 272–288.

Lakoff, G. (1987). *Women, Fire and Dangerous Things*: *What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Liberman, A.M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how, much they know? *Organizational Behavior and Human Performance*, 20, 159-183.

Lichtenstein, S., Fischhoff, B., & Phillips, S. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306-344). Cambridge University Press.

Link, S. W. (1992). *The Wave Theory of Difference and Similarity*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492-527.

Love, B. C. (2002).Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, 9,829-835.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 671-676). Cambridge, MA: MIT Press.

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

Maddox, W. T., & Ashby, F. G. (1993).Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, 53, 49-70.

Maddox, W.T., & Ashby, F.G. (2004) Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes*, 66, 309-332.

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, & Cognition*, 29, 650-662.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition*, 32, 582-591.

Maddox, W. T., &Ing, A.D. (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, 100-107.

Maddox, W.T., Filoteo, J.V, Hejl, K.D., & Ing, A.D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 227-235.

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, 108, 578–589.

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

Mandler, J. M. (1988). How to build a baby: On the development of an accessible representationalsystem. *Cognitive Development*, 3, 113-136.

McKinley, S. C.,& Nosofsky, R. M. (1995). Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 128-148.

McKinley, S. C., & Nosofsky, R. M., (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 294-317.

Medin, D. L., & Atran, S. (Eds.), (1999). *Folkbiology*. Cambridge: MIT Press.

Medin, D. L., Lynch, E. B. & Solomon, K. E. (2000). Are there kinds of concepts. *Annual Review of Psychology,* 51, 121-147.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 355–368.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19, 242-279.

Merikle, P. M., & Reingold, E. M. (1991). Comparing direct (explicit) and indirect (implicit) measures to study unconscious memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*,17, 224–233.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the *AB* task. *Developmental Science*, 1, 161-184.

Munakata, Y., Morton, J. B., & Stedron, J. M. (2001). The role of prefrontal cortex in preservation: developmental and computation explorations. In Quinlan P. (ed.), *Connectionist Models of Development*. Hove: Psychology Press.

Murphy, G. (2002). *The Big Book of Concepts*. London: MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20, 904-919.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.

Nelson, T. O., & Narens L. (1990). Metamemory: A theoretical framework and new findings. In G H Bower (Ed) *The Psychology of Learning and Motivation* (Vol 26 pp 125-173) New York Academic Press.

Newell, B. R., Dunn, J. C.,& Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, *38, 563-581.*

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirement of learning: evidence from performance measures. *Cognitive Psychology*, 19, 1-32

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

Norman, G. R., Brooks, L. R., Coblentz, C. K., & Babcook, C. J. (1992). The correlation of feature identification and category judgments in diagnostic radiology. *Memory & Cognition,* 4, 344–355.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latencyin retrieval from short-term memory. *Journal of Mathematical Psychology*, 6, 192-208.

Norman, D. A., & Shallice, T. (1980). *Attention to Action: Willed and Automatic Control of Behavior*. University of California, San Diego, CHIP Report 99.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 10, 104-114.

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

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 15, 282-304.

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.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of multiple-system phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375-402.

Nosfosky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychonomic Bulletin & Review,* 9 169-174.

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar- based interpretation. *Psychological Science*, 9, 247-255.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cerebral Cortex*, 12, 247-257.

Pallier, G., Wilkinson, R., Danthiir, V., & Kleitman, S. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology,* 129, 257–299.

Paul, E. J., Boomer, J., Smith, J. D., & Ashby, F. G. (2011). Information–integration category learning and the human uncertainty response. *Memory Cognition*, 39, 536–554.

Pierce, C. S., & Jastrow, J. (1884). On small differences in sensation. *Proceedings of the National Academy of Sciences*, 3, 75-83.

Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision,* 10*,* 437-442.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature, 331,* 985-989.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 962-986.

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in perceptual judgments. *The American Journal of Psychology*, 110, 543-572.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin and Review*, 10, 177-183.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15, 285-290.

Pisoni, D. B., Aslin, R. N., Percy, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 297-314.

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

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia*, 41, 245-251.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 304–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304-308.

Poulin-Dubois, D. (1999). Infants' distinction between animate and inanimate objects: The origins of naive psychology. In P. Rochat (Ed.), *Early Social Cognition: Understanding Others in the First Months of Life* (pp.257-280). Hillsdale, NJ: Erlbaum.

Rabbitt, P. (1966a). Error correction time without external signals. *Nature*, 212, 438.

Rabbitt, P. (1966b). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, 71, 264–272.

Rajaram, S., Hamilton, M., & Bolton, A. (2002). Distinguishing states of awareness from confidence during retrieval: Evidence from amnesia. *Cognitive, Affective, and Behavioural Neuroscience*, 2, 227-235.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior, 6,* 855-863.

Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. New York: Oxford.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 382- 407.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 811–829.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27, 1261-1275.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122,* 92-114.

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, 44, 563–575.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosnaidou & A. Ortony (Eds.) *Similarity and Analogical Reasoning*. Cambridge University Press: New York.

Roberts, A. C., Robbins, T. W., & Everitt, B. J. (1988).The effects of intradimensional and extradimensional shifts on visual discrimination learning in humans and non-human primates.*The Quarterly Journal of Experimental Psychology Section B*, 40, 321 - 341.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129, 369-398.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*,4, 328–50.

Rosch, E. (1975). The nature of mental codes for color categories. *Journal at Experimental Psychology: Human Perception and Performance*, 1, 303-322.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology,* 7, 532–47.

Rosch, E., & Mervis, C. B.(1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.

Rosenhan, D. (1973). On being sane in insane places. *Science*, 79, 250-252

Scheck, P., & Nelson, T. (2005). Lack of pervasiveness of the underconfidence-with-practice-effect; boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*, 134, 124–128.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*. 84, 1-66.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools Inc.

Schoenherr, J. R. (2008a). *The Dependence of Confidence Processing on Working Memory* (Unpublished Master's thesis). Carleton University, Ottawa.

Schoenherr, J. R. (2008b). *Confidence Processing as an Integrated Metamonitoring Operation*. Unpublished manuscript.

Schoenherr, J. R. & Petrusic, W. M. (2011). *Scaling Internal Representations of Confidence: Effects of 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.

Schoenherr, J. R., & Logan, J. (2013). *Attending Unattended Regions of an Acoustic Continuum: Modulation of Acoustic and Phonemic Representations in Speech Perception by Feedback*. Unpublished Manuscript.

Schoenherr, J. R. & Logan, J. (2014). Attentional and immediate memory capacity limitations in the acquisition of non-native linguistic contrasts. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, Québec City, Canada.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feeling of knowing. *Psychonomic Bulletin and Review,* 1, 357-375.

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *Behaviour Brain Science,* 17, 367–447.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75, (13,*Whole No. 517).

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127-190.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, 121, 278-304.

Shynkaruk, J., M. & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, 34, 619-632.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 23, 659-680.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411-1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 3-27.

Smith, E. E., Patalano, A. L. & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, 65, 167-196.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago, IL: University of Chicago Press.

Stanovich, K. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645–726.

Stewart, N.,& Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 893–907.

Sumner, F. B. (1898). A statistical study of belief. *Psychological Review*, 5, 616-631.

Szmalec, A., & Vandierendonck, A. (2007).Estimating the executive demands of a one-back choice reaction time task by means of the selective interference paradigm. *Quarterly Journal of Experimental Psychology*, 60, 1116–1139.

Szmalec, A., Vandeirendonck, A., & Kemps, E. (2005). Response selection involves executive control: Evidence from the selective interference paradigm. *Memory & Cognition*, 33, 531-541.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

Taylor, S. F., Stern, E. R. & Gehring, W. J. (2007). Neural Systems for Error Monitoring: Recent Findings and Theoretical Perspectives. *Neuroscientist,* 13; 160-172.

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 10, 119-139.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition*, 31, 1060-1071.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108, 550-592.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, 2, 442-459.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582-600.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50, 179-197.

Vosnaidou, S., & Ortony, A. (1989). *Similarity and Analogical Reasoning*. Cambridge University Press: New York.

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, 168-176.

Wixted, J., T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review,* 11, 616-641.

Werker, J. F., & Logan, J. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37, 35-44.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Yaniv, I., Yates, J. F. & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, 110, 611-617.

Yonelinas, A. P.(2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46, 441–517.

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

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

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Science*, 7, 214–218.

Zelazo, P. D., Moscovitch, M., & Thompson, E. (2007). *The Cambridge Handbook of Consciousness*. New York: Cambridge University Press

Ziori, E., & Dienes, Z. (2006). Subjective measures of unconscious knowledge of concepts. *Mind & Society*, 5, 105–122.

Ziori, E., & Dienes, Z. (2008). How does prior knowledge affect implicit and explicit concept learning? *The Quarterly Journal of Experimental Psychology*, 61, 601-624.

Appendix: Computing Confidence Calibration Indices

Contemporary studies of confidence generally rely on numeric labels using numbers that increase in magnitude (e.g., 1 through 7; Vickers, 1979), proportions (.0 through 1.0) or percent correct (0% through 100%). The use of numeric labels affords the researcher an opportunity to compare the performance obtained in the primary task to the subjective reports (i.e., Budescu & Wallsten, 1990). An additional benefit of using values that correspond to the objective probability of an event's occurrence is that it allows for a direct comparison between confidence ratings and performance in a task. A number of quantitative techniques have been proposed. In addition to examining mean confidence across conditions, the present study used two confidence calibration indices: calibration and over/underconfidence (e.g., Barasnki & Petrusic, 1994). They were selected because they are widely used and readily interpretable measures of the relationship between task performance and confidence judgments (e.g., Keren, 1991; for comprehensive discussions of a number of measures, see Zelazo et al., 2007). Moreover, they will allow me to interpret the thesis's confidence results in the light of this established literature.

A measure of *calibration* allows researchers to examine the extent to which the subjective probability assigned by participants to an event corresponds to the objective probability of the event. For instance, when presented with a domesticated animal, one might use the diagnostic feature "striped fur coat" to assign it to either the category of *Dog* or *Cat*. With two categories, the subjective probability that a participant is correct would range from guess (0.5) to certainty (1.0). The strong association of *Cat* with stripes given our experience would cause us to assign a striped pet to the *Cat* category and assign the event a high degree of confidence (e.g., 0.9). The extent to which this subjective probability corresponds to an objective

probability (e.g., Striped cats might only occur with a probability of 0.8 in the world) might be translated into a measure known as calibration.

In order to compute calibration, the squared difference between the *j*th subjective probability category, ψ_j, and the proportion correct for events, $p(cor)_j$ must be calculated. The product of these responses over the number of trials, n_j, on which they were used is summed across all subjective probability categories (J) and a mean calibration score is obtained. Formally, this is given by the equation (Lichtenstein & Fischhoff, 1977):

$$\bar{\quad}$$

Calibration then varies between 0.0 (there is no difference between the objective occurrence of an event and subjective assessment of performance) and 1.0 (there is no correspondence between objective and subjective assessment of performance). Calibration thus represents the unsigned difference between obtained and subjective performance. Baranski and Petrusic (1994) note that calibration is rarely worse than 0.10. This could be taken as suggesting that in most perceptual discrimination and judgment tasks, participants exhibit some awareness of their performance.

Whereas calibration establishes the extent to which participants' confidence judgments track their actual performance on a given task, over/underconfidence measures whether participants' beliefs overestimate or underestimate their actual performance. The computation of overconfidence is comparatively straightforward relative to calibration. A researcher obtains the

difference between the mean subjective probability provided by a participant for any given condition, ψ_j, and the proportion correct for events, $p(cor)_j$. The equation is simply:

$$OC = \quad - \quad _j$$

Unlike calibration, overconfidence represents the signed difference between mean performance and subjective probability for any condition. Negative values represent underestimation of performance, or underconfidence, whereas positive values represent overestimations of performance, or overconfidence. An index of overconfidence thereby provides information about the overall awareness of a participant rather than simply whether their performance was above chance when using the guessing response (i.e., underconfident in the 50% category).

It is important to note that notwithstanding their intimate relationship, calibration and overconfidence do not reveal the same information about participants' subjective awareness. Poor calibration is not always the result of a systematic bias in the participants' subjective assessment of their performance. For instance, participants could be overconfident when using small confidence values (e.g., report 60% confidence when only when 50% accurate) and underconfident when using larger confidence values (e.g., report 70% when they are 90% accurate). The aggregate of this pattern could potentially yield no systematic pattern of overconfidence. In short, overconfidence bias requires miscalibration: as overconfidence increases so does subjective miscalibration. Miscalibration, however, need not be reflected in a tendency to be overconfident: participants' subjective miscalibration could increase while yielding overestimations on some trials and underestimation of on others.