ORIGINAL ARTICLE

Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning

C. E. R. Edmunds[1] | **Fraser Milton**[2] | **Andy J. Wills**[1]

[1]School of Psychology, Plymouth University, Plymouth, Devon, PL4 8AA, UK

[2]School of Psychology, University of Exeter, Exeter, Devon, EX4 4QG, UK

Correspondence
C. E. R. Edmunds, School of Psychology, Plymouth University, Plymouth, Devon, PL4 8AA, UK
Email: ceredmunds@gmail.com

Funding information

Behavioral evidence for the COVIS dual-process model of category learning has been widely reported in over a hundred publications (Ashby and Valentin, 2016). It is generally accepted that the validity of such evidence depends on the accurate identification of individual participants' categorization strategies, a task that usually falls to Decision Bound analysis (Maddox and Ashby, 1993). Here, we examine the accuracy of this analysis in a series of model-recovery simulations. In Simulation 1, over a third of simulated participants using an Explicit (conjunctive) strategy were misidentified as using a Procedural strategy. In Simulation 2, nearly all simulated participants using a Procedural strategy were misidentified as using an Explicit strategy. In Simulation 3, we re-examined a recently-reported COVIS-supporting dissociation (Smith et al., 2014), and found that these misidentification errors permit an alternative, single-process, explanation of the results. Implications for due process in the future evaluation of dual-process theories, including recommendations for future practice, are discussed.

KEYWORDS
COVIS, Decision Bound analysis, dual-systems, categorization strategies

1 | INTRODUCTION

There is substantial evidence, and a degree of consensus, that different participants in the same categorization task can use a range of qualitatively different strategies (Ashby et al., 1998; Meeter et al., 2006; Nosofsky and Zaki, 2002; Raijmakers et al., 2001; Wills et al., 2015). Nevertheless, some researchers compare averaged measures of learning performance between category structures or task conditions in order to draw conclusions about the likely underlying mechanisms of category learning. To draw these inferences validly, the participants in each condition must all learn in a qualitatively similar way (Estes, 1956; Kurtz, 2015; Sidman, 1952). If this is not the case, the average will likely not represent the behavior of any single person, causing severe interpretative difficulties (Lee and Webb, 2005; Maddox, 1999; Navarro et al., 2006; Siegler, 1987).

COVIS (COmpetition between Verbal and Implicit Systems; Ashby et al., 1998, 2011) is one category learning model that aims to predict when and why participants use different strategies. COVIS assumes that categorization

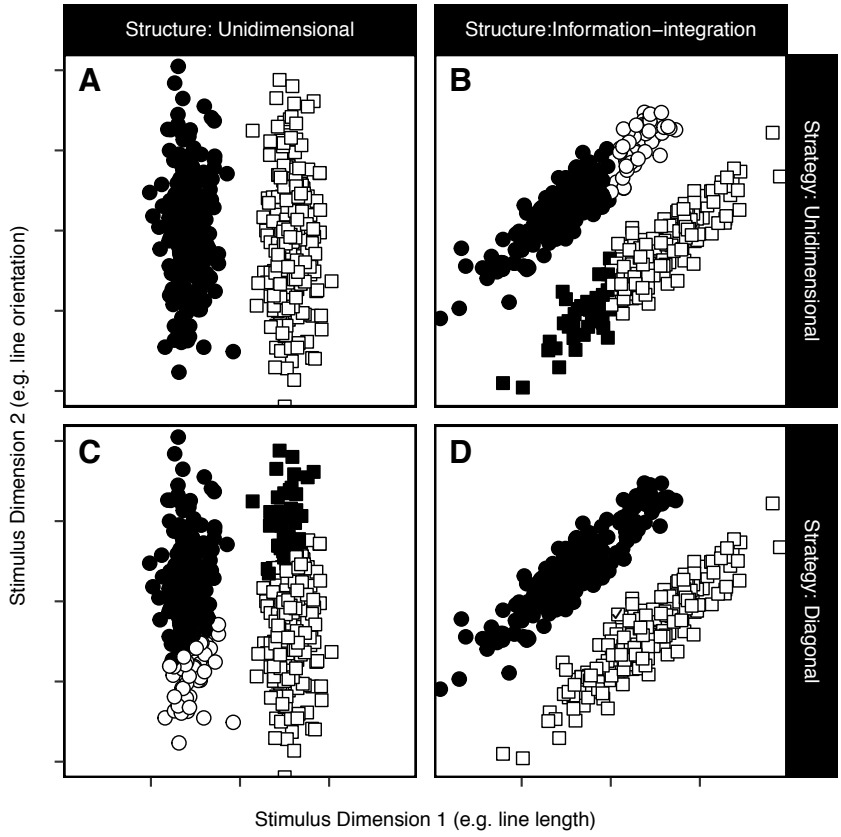

FIGURE 1 Two example strategies implemented as linear decision boundaries through hypothetical two-dimensional stimulus spaces representing unidimensional and information-integration category structures. Each point represents a stimulus. Circles represent those stimuli defined as being in Category A, and squares those defined as Category B. Filled shaped indicate those stimuli assigned to Category A by the strategy, and unfilled those to Category B.

performance is mediated by two, parallel, competing systems: an Explicit System and a Procedural System. The Explicit System is assumed to implement rule-based strategies using either one stimulus dimension (i.e., unidimensional rules such as in Fig. 1A; Smith et al., 2014; Ashby and Valentin, 2016) or rules based on easily verbalizable combinations of stimulus dimensions (i.e., conjunction rules; Casale et al., 2012; Filoteo et al., 2010; Zeithamova and Maddox, 2006). Because the Explicit System implements these rule-based strategies, COVIS predicts that this system will optimally learn rule-based category structures, those with category boundaries lying either parallel or perpendicular to the stimulus dimensions (such as in Fig. 1A and C or conjunctions).

If rule-based strategies do not result in high enough accuracy–because the category structure is not rule-based and thus difficult to verbalize–COVIS predicts that the Procedural System will gain control of responding. As the Procedural System is predicted to implement a variety of strategies (including the one demonstrated in Fig. 1C and D) it can implement the optimum strategy for information-integration category structures, where stimulus dimensions need to be combined (as shown in Fig. 1D).

Experiments used to support the COVIS model typically compare the effect of an experimental manipulation (e.g. concurrent load) on the learning of rule-based and information-integration category structures (see Ashby and Maddox, 2011; Ashby and Valentin, 2016, for some recent reviews). By varying the category structures participants learn, the experimenters hope to elicit a switch in the learning system participants use to control responding. In other words, typical COVIS-supporting experiments are based on the assumption that because participants are learning a rule-based or information-integration category structure they will learn to use the appropriate rule-based or procedural strategy and therefore be using the Explicit or Procedural System, respectively. Then, if the experimental manipulation affects learning of one category structure more than another, the experimenter infers that it affects one system more than the other, thereby providing evidence for a dual-system model of category learning (although see Newell et al., 2011, for a discussion of the limitations to this approach).

For this type of experiment, the presence of subsets of qualitatively-different participants can be particularly problematic (for an example in the literature, see Newell et al., 2010). Critically, the conclusion that the experiment supports a dual-system model depends on the assumption that the participants in each category structure condition used the most appropriate system to learn those structures. Specifically, that participants used the Explicit System to learn the rule-based category structure, and the Procedural System to learn the information-integration category structure. If this is not the case, then the introduction of subsets of participants using sub-optimum strategies may result in inferential errors.

To illustrate this point, consider an experiment that examined the effect of concurrent load and found it caused a reduction in performance by 10% for a particular category structure (such as in Zeithamova and Maddox, 2006). In the ideal case, all participants classifying a particular category structure would be using the same optimum strategy and all those in the relevant condition would be similarly affected by the manipulation; participants with concurrent load would score around 10% less than they would have without the load. Here, we could use standard group-accuracy analyses validly.

However, if some participants use other, sub-optimum strategies then drawing valid conclusions from this experiment is much harder. One possibility is that the manipulation, within a given category structure, changes the relative proportions of different strategies used in each condition (load vs. no load). This would result in a change in average accuracy because, given a particular category structure, the highest performance possible for each strategy varies. A second possibility is that the manipulation did not cause a change of strategies, but rather had a differential effect depending on the strategy type being used. For example, the manipulation could have had no effect on people using the optimum strategy, but could severely affect performance reliant on sub-optimum strategies (see Schnyer et al., 2009, for a similar argument). Indeed, when Newell et al. (2010) repeated Zeithamova and Maddox's experiment, they removed

participants below a learning criterion and found that the effect vanished. The experiment was no longer consistent with a dual-systems account of category learning.

To avoid the possibility that any dissociations in accuracy are due to different proportions of sub-optimum strategies, COVIS-supporting experiments typically include a decision-bound strategy analysis (hereafter, *DB analysis*). DB analysis is a special case of General Recognition Theory (GRT; Ashby and Gott, 1988; Ashby and Soto, 2015), which is a multidimensional generalization of signal detection theory (for a further discussion of the relationship between DB analysis and GRT; see Helie et al., 2017). Although decision-bound models have a number of different uses (see e.g. Fific et al., 2010), COVIS-supprorting experiments use them as a manipulation check to determine which strategy each participant is using. This approach assumes that strategies can be modeled by a (usually linear) decision bound that passes through stimulus space (such as those in Fig. 1). For each participant, a variety of DB models are fitted to their responses and the one that best represents that participant's pattern of responding is selected. To determine whether the manipulation check was passed, each participant's strategy is compared to the category structure they were assigned to learn. If a sufficient number of participants are found to be using the optimum strategy for the category structure they were assigned, then the category structure manipulation is assumed to have elicited a corresponding shift in category learning system. Under this assumption, any dissociations in accuracy are then ascribed to the existence of dual systems (Ashby and Maddox, 2005, 2011; Ashby and Valentin, 2016).

Using DB analysis (or any other analysis method) as a manipulation check is logically valid if and only if that analysis method consistently and accurately identifies the strategies that participants are using; in the case of DB analysis, it must be able to correctly identify strategies under a variety of differing category structures, experimental manipulations and levels of noise. However, recent evidence from our lab suggests that DB analysis does not accurately recover the strategies participants use for information-integration category structures. Edmunds et al. (2015) re-examined an experiment by Ashby et al. (2002) that examined the effect of feedback type on learning of rule-based and information-integration category structures. However, Edmunds et al. extended the original protocol by asking participants to describe their response strategies. Crucially, a substantial number of responders classified as using the Procedural System on the basis of DB analysis nevertheless reported using an explicit rule-based strategy.

One possible explanation for this contradiction is that participants did not accurately report their strategies. There are two pieces of evidence that speak against this interpretation. First, verbal reports successfully predict participants' performance in other category learning tasks (Lagnado et al., 2006). Second, previous evidence has demonstrated that verbal reports lack sensitivity and thus underestimate the prevalence of explicit processes (e.g., Shanks and St John, 1994). Therefore, the large number of explicit reports given in Edmunds et al. (2015) would, if anything, underestimate the number of participants using explicit sorting strategies.

An alternative, more problematic, explanation for the disparity between the strategies identified by verbal report and by DB analysis in Edmunds et al. (2015) is that DB analysis is biased toward finding the optimal strategy for the presented category structure (i.e. towards finding rule-based strategies for rule-based category structures and implicit strategies for information-integration category structures). One possible conjecture is that DB analysis has this bias, at least in part, because it typically uses just the training stimuli, rather than a broad range of transfer stimuli, to determine the participants' strategy. Work by Donkin et al. (2015) provides some support for this conjecture. Donkin et al. compared the results of a DB analysis conducted on just the training stimuli with a DB analysis conducted on both the training and transfer stimuli. The addition of the transfer stimuli resulted in stimuli being more evenly distributed across the stimulus space. They found that when they added in the transfer stimuli, the proportion of participants classified as using the optimal (diagonal) strategy for an information-integration category structure fell. At the very least, this suggests that the conclusions we can draw from the DB analysis depend on the distribution of stimuli across stimulus space.

To summarize, there is some evidence that DB analysis does not accurately recover the strategies participants use. This makes determining whether category learning is mediated by two competing learning systems much more difficult. For instance, consider an experiment that found that feedback delay harmed information-integration category learning but had no effect on unidimensional rule-based category learning (such as Ell et al., 2009, although see Dunn et al., 2012). Furthermore, suppose that DB analysis found that all the participants used the optimum strategy for the category structure they were presented with. If DB analysis was accurate, we might conclude that the source of this interaction was the presence of two different systems. However, if DB analysis was inaccurate this inference would not be the only one we could make. For example, if DB analysis, in the information-integration conditions, falsely identified an explicit conjunction rule strategy as a diagonal (procedural) strategy, an alternative account might be that feedback delay impacts learning once participants are using sufficiently complex rules. This would be consistent with a single-system account (related arguments have been made in Edmunds et al., 2015; Nosofsky et al., 2005). Another possibility is that we might incorrectly reject a dual-system account of category learning, because an unreliable DB analysis failed to show a difference in strategies used between conditions.

However, a limitation of virtually all work to date is that one can never be sure whether DB analysis contains significant flaws, because one does not know which strategy participants were actually using. When employing data from real participants, all we have are multiple forms of assessment of their strategy (DB analysis, verbal reports, brain activation…), all of which provide indirect and potentially flawed information. Using one measure to assess the quality of the others introduces the circularity of assuming one of the measures is correct. In the current article, we use a model-recovery approach to break out of this loop.

Model recovery involves simulating hypothetical participants' responses according to the strategy models defined by the strategy analysis. By simulating responses, we circumvent the problems with Edmunds et al. (2015) as now we know exactly which model each (simulated) participant is using. From these hypothetical, simulated participants we can then use DB analysis to identify the strategies from the responses to see whether DB analysis is capable of recovering the correct generating model. This model-recovery procedure is recommended as best practice for any cognitive modeling analysis (Heathcote et al., 2014) but has yet to be done for the typical DB analysis used in the COVIS literature. In fact, to our knowledge, there is just one previous report of a model-recovery simulation in this area, and it was conducted for purposes somewhat different to our own. Specifically, Donkin et al. (2015) demonstrated that some models not normally used in DB analysis (highly complex rules) were seldom mis-recovered from simulated participants doing something not normally considered in DB analysis (prototype-based responding), in an experimental design that reduced the likelihood of mis-recovery relative to more standard experiments in the COVIS literature (a point we will return to in the General Discussion).

Another potential limitation of the work by Edmunds et al. (2015) and Donkin et al. (2015) relates to their choice of category structures. Donkin et al. (2015) used a four-category information-integration category structure, and Edmunds et al. (2015) included a conjunction rule-based category structure. However, in the COVIS literature, arguably the most common category structures are the two-category unidimensional and information-integration category structures shown in Fig. 1. These category structures are often considered within the COVIS literature to be good choices for comparing explicit and procedural category learning (although see Carpenter et al., 2016; Edmunds et al., 2015; Nosofsky et al., 2005, for arguments to the contrary) because they differ markedly in verbalizability whilst being matched on several key attributes such as within-category similarity, between-category distance and the optimal accuracy a participant could achieve (usually 95% or above; Smith et al., 2014, 2015).

In our first two model-recovery simulations, we consider the effectiveness of DB analysis in relation to these information-integration (Simulation 1) and unidimensional (Simulation 2) category structures. We then provide a specific demonstration (Simulation 3) of how the issues so revealed permit a major reinterpretation of a published experiment

(Smith et al., 2014). We conclude with some recommendations for future work. Some of these recommendations derive directly from our simulations, others are more generally just good practice. We hope these recommendations may lessen the future impact of the issues revealed by our simulations.

2 | SIMULATION 1: INFORMATION-INTEGRATION CATEGORY STRUCTURE

2.1 | Method

The details of the model-recovery procedure are briefly described here; for more extensive details describing all modeling procedures see the Appendix.

First, we generated a two-dimensional information-integration category structure using the procedure outlined by Ashby and Gott (1988) according to the parameters reported by Smith et al. (2014). For this structure, each category i is generated by sampling points from a bivariate normal distribution with mean μ_i and covariance matrix Σ_i. Each point represents a stimulus, with the x-value corresponding to one stimulus dimension and the y-value corresponding to the other stimulus dimension. Then, we determined the optimal unidimensional, conjunction and diagonal (GLC) strategies for that information-integration category structure. We then generated the responses of the simulated participants by adding various levels of noise to each of these three strategies (simulating 20 participants for each strategy and level of noise). Finally, we conducted a DB analysis on each simulated participant.

The DB analysis used in the COVIS literature aims to find, for each participant, the decision bound through stimulus space that best separates their responses for one category from the other[1]. Different strategy types are implemented by varying the functional form of the decision boundary. For example, a unidimensional strategy is implemented by a straight line perpendicular to one of the stimulus dimensions.

DB analysis assumes that stimulus perception is subject to normally distributed noise: each time a participant sees a particular stimulus it is perceived slightly differently (Ashby and Gott, 1988). Therefore, a stimulus near the decision boundary is more likely to be misclassified because small amounts of noise may make it appear to the participant as if that stimulus was on the wrong side of the boundary. Each participant's strategy is determined by fitting multiple decision-bound models, with different functional forms, to their response data and some measure of fit is calculated for each model. The functional form of the model that best describes that participant's pattern of responding gives their strategy.

The two kinds of model-based strategies that are of particular interest within the COVIS literature are rule-based and information-integration strategies (Ashby and Gott, 1988; Maddox and Ashby, 1993). This is because these kinds of strategy type are hypothesized to be implemented by the Explicit and Procedural Systems of COVIS respectively (Ashby et al., 1998, 2011). Within the COVIS literature, rule-based strategy models are implemented by linear decision boundaries that are parallel or perpendicular to the stimulus dimensions in stimulus space (Maddox and Ashby, 1993). Typical examples include unidimensional and conjunction rules. A unidimensional rule consists of a single decision boundary orthogonal to the relevant dimension. It has up to two parameters: perceptual variance and the value of the boundary on the relevant dimension. For example, a unidimensional rule on basis of line length can be verbalized as "If the line is short it is in Category A, if it is long it is in Category B." A conjunction rule consists of two decision boundaries orthogonal to each other. It has up to four parameters: perceptual variance in two dimensions and the values where the two boundaries cross the axes. For example, this might correspond to the rule "If the line is short and upright it is in

[1] For ease of explanation we have focused on the case where the stimuli under consideration have two stimulus dimensions and are being sorted into two categories. This is not always the case (cf. Maddox et al., 2004c), although it is the most frequent instantiation of this theory and analysis within the COVIS literature.

Category A; otherwise Category B." These strategies are assumed to be implemented by the Explicit System and to be optimum for rule-based category structures.

The information-integration strategy model is usually implemented in the COVIS literature by a diagonal strategy, also known as the General Linear Classifier (GLC; Maddox and Ashby, 1993). The diagonal strategy consists of a single linear decision boundary in stimulus space which is not parallel to any of the stimulus dimensions. This strategy has up to three parameters: perceptual variance, the slope and y-intercept of the line. This strategy is generally considered difficult to verbalize and so is hypothesized to be implemented by the Procedural System and to be the optimum strategy for learning information-integration category structures (Ashby et al., 1998, 2011).

In addition to the sets of models that correspond to the Explicit and Procedural Systems of COVIS, researchers within the COVIS literature often also include random models. Random models do not assume the existence of a decision boundary (Maddox and Ashby, 1993). Rather they assume that stimulus features are irrelevant to responding and that participants respond at random. Two types of random model are usually included: one with no parameters that assumes that participants respond equally to both categories, and one with one parameter which represents biased responding toward one category.

The best fitting model is selected from the models above for each participant using a measure of model fit. Here, we have chosen to define the best strategy model as the one that minimises the Bayesian Information Criterion (BIC; Schwarz, 1978) as this is the measure of fit most often used in the COVIS literature (for instance, Smith et al., 2014).

Once the best-fitting strategy model has been selected from the models above for each participant using a measure of fit, proponents of COVIS use these strategies to check the experimental assumptions of COVIS (Ashby and Maddox, 2005, 2011; Ashby and Valentin, 2016). Typically, they look at the proportion of each strategy type in each cell of the experimental design. For any dissociation in accuracy scores to support the COVIS model of category learning, there must be a corresponding difference in the types of strategy identified across the category structure conditions. If this is the case, researchers assume that their category structure manipulation was successful in inducing an equivalent change in category learning system. Under this assumption, any dissociation in the effect of a manipulation across category structures can be attributed to the manipulation differentially affecting two underlying learning systems.

2.2 | Results and Discussion

Table 1 shows the proportion of simulated participants who were identified as using the unidimensional, conjunction, diagonal (GLC) and random strategies, as a function of their generating strategy (for a visual representation of this information, please see the Supplementary Material associated with this article). Also shown are the Schwarz weights for the winning strategy models (Wagenmakers and Farrell, 2004)[2]. Schwarz weights ($wBIC$) are defined as the probability that a particular model i is best, in terms of minimising the BIC, given the data and the set of competing models j.

The diagonal generating strategy (GLC) is recovered well, as is the unidimensional generating strategy (UD). Occasionally, they were mis-identified as a random strategy. This kind of misidentification has little theoretical impact.

In contrast, the mis-identification of the conjunction generating strategy (CJ) is substantively problematic. Simulated participants whose responses were generated using a conjunction strategy were more likely to be incorrectly recovered than correctly recovered, with a diagonal (GLC) strategy as the most likely mis-recovery.

This finding raises a wide-ranging uncertainty about the existing COVIS literature, as it indicates that the proportion of participants using a rule-based strategy to learn an information-integration category structure may be much higher

[2]Note that here, unlike other applications of Schwarz weights, the values would not be expected to sum to one across rows or columns, as they are reported for the winning strategy models only.

TABLE 1 The proportion of each type of strategy recovered for each type of generating strategy for an information-integration category structure.

Generating strategy	Recovered strategies *(wBIC)*			
	UD	**CJ**	**GLC**	**RND**
UD	0.99 *(0.97)*	-	-	0.01 *(0.00)*
CJ	0.14 *(0.10)*	0.48 *(0.46)*	0.38 *(0.26)*	-
GLC	0.05 *(0.03)*	0.06 *(0.05)*	0.83 *(0.76)*	0.06 *(0.05)*

Strategies: UD=Unidimensional, CJ=Conjunction, GLC=General linear classifier (diagonal), RND=Random.

than previously estimated. One consequence of a high proportion of rule-based participants being classified as using a Procedural strategy is that some studies are likely to have provided false evidence for the existence of two learning systems (Type I error). This is because this type of mis-identification would make it appear that the category structure manipulation was successful when it was not: participants were still using rule-based strategies to learn an information-integration structure. Then, researchers might incorrectly infer that a manipulation designed to impair procedural learning had successfully done so. However, the drop in performance would more likely be due to participants using more complex, rule-based strategies in the information-integration category structure condition than in the rule-based category structure. Therefore, this would be consistent with a single-system rule-based account (as is demonstrated in Simulation 3).

3 | SIMULATION 2: UNIDIMENSIONAL CATEGORY STRUCTURE

Recall that COVIS makes predictions about the strategies that cause a participant's responses, while DB analysis works solely on the responses they actually produce. This means that for DB analysis to be useful in the assessment of the COVIS model, the different responses people make must be able to discriminate between the different strategies they are using.

Unfortunately, with the unidimensional category structures typically used in the COVIS literature (e.g., Smith et al., 2014, 2015), it can be difficult to discriminate between different underlying strategies. This is because it is possible to achieve 100% accuracy on a unidimensional category structure using strategies other than the optimal unidimensional one. For example, a conjunction strategy can also score 100% on a unidimensional category structure if one bound passes between the two categories and the other bound passes either above or below the categories. A diagonal (GLC) strategy can also score 100% on a unidimensional category structure if the bound is steep enough so as to pass between the two categories.

Intuitively, it seems likely that DB analysis would identify all of these different underlying strategies as unidimensional. This is because DB analysis uses measures of model fit that penalize the number of parameters in each model (Akaike, 1974; Schwarz, 1978). Therefore, if a participant used a diagonal (GLC) strategy on a unidimensional category structure, it is likely that DB analysis would mis-identify that participant as using the optimum unidimensional strategy, because the unidimensional DB model has two parameters whereas the diagonal (GLC) model has three. As discussed by Wills and Pothos (2012), one of the problems of this kind of "Occam's Razor" model selection is that a poor selection of evidence can lead to an overly-simple model being favored.

If our intuitions were correct, this could be problematic for the existing COVIS literature. COVIS predicts that

TABLE 2 The proportion of each type of strategy recovered for each type of generating strategy for a unidimensional category structure.

Generating strategy	Recovered strategies *(wBIC)*			
	UD	**CJ**	**GLC**	**RND**
UD	0.97 *(0.95)*	-	-	0.02 *(0.02)*
CJ	0.95 *(0.88)*	0.04 *(0.03)*	-	0.01 *(0.01)*
GLC	0.98 *(0.95)*	-	-	0.02 *(0.02)*

Strategies: UD=Unidimensional, CJ=Conjunction, GLC=General linear classifier (diagonal), RND=Random.

participants should be using a verbalizable rule-based strategy when learning a unidimensional category structure. However, if one cannot determine from the standard DB analysis whether a participant identified as using a unidimensional strategy is actually doing so, then DB analysis cannot be used to test this hypothesis.

3.1 | Method

To test this intuition we conducted a simulation similar to the one in the previous section (for more details, see the Appendix). First we generated a unidimensional category structure using the procedure outlined by Ashby and Gott (1988), similar to that used in Smith et al. (2014, 2015). Then, we fitted a unidimensional, conjunction and diagonal (GLC) strategy to this category structure to determine the optimum strategy of that type for this category structure. For example, the optimum diagonal (GLC) strategy for this category structure is a steep diagonal line that passes between the two categories.[3] We then simulated the responses of hypothetical participants according to this optimum strategy. To these responses, we added additional noise. Finally, we conducted DB analysis on each simulated participant and calculated the proportion of participants who were identified as using the unidimensional, conjunction, diagonal (GLC) and random strategies.

3.2 | Results and Discussion

The results from this simulation are shown in Table 2 (for a visual representation of this information, please see the Supplementary Material associated with this article). The simulation confirms our intuition: DB analysis cannot discriminate well between strategies applied to the unidimensional category structure.

Note first that DB analysis does do well when the generating strategy is unidimensional (the first row of Table 2). Here, the recovered strategy models are either the correct unidimensional strategy or (occasionally) the random model. However, if we take this as a marker of good performance, we can see that the more complex models are recovered much less well. For example, just 4% of the simulated participants who were using a conjunction strategy were correctly identified as doing so, whereas 95% of them were mis-identified as using a unidimensional strategy. Similarly, none of the simulated participants who used the diagonal (GLC) strategy were correctly identified as doing so. In contrast, 98% of them were mis-identified as using a unidimensional rule-based strategy.

[3] Of course, analytically a vertical line is a special case of the diagonal (GLC) strategy. However, in practice, the model fitting procedure still attempts to estimate a gradient for the line. Then, because the model is fitted using an algorithm that uses gradient descent, there comes a point were even a large increase in the gradient of the line results in little to no improvement in fit. Thus, in practice, the optimum diagonal (GLC) model is close to vertical, but still diagonal.

From the perspective of the COVIS model, misidentifying a diagonal (GLC) strategy as unidimensional is much more problematic than misidentifying a conjunction strategy as unidimensional. This is because the diagonal strategy is assumed to be a marker for the Procedural System while unidimensional and conjunction strategies are assumed to be a marker for the Explicit System (Filoteo et al., 2010; Zeithamova and Maddox, 2006). Thus, if a participant were to be learning a unidimensional category structure procedurally, DB analysis would misleadingly indicate that they were learning it explicitly.

Of course, the hypothesis that participants learn unidimensional category structures non-explicitly can sometimes be discounted by other means. For example, participants' verbal reports of using a unidimensional strategy predict their classification behavior in some procedures (e.g. Wills et al., 2013). More generally, one might argue that: (1) it seems unlikely that participants would use diagonal (GLC) strategies to learn unidimensional category structures as it would require attention to more stimulus dimensions and, therefore that, (2) in practice, misidentification of strategies for unidimensional category structures is unlikely. One might then conclude that, (3) the particular failing of DB analysis revealed by Simulation 2 may not have much theoretical impact. In contrast, we argue that a modeling approach (such as DB analysis), which aims to represent participants' responses accurately enough to be a manipulation check should be unbiased. It should work well across the board (i.e. for different category structures) not just in those situations our current theories deem likely. Additionally, DB analysis is also used in experiments where participants are trained on one set of stimuli, and tested on others (e.g., Edmunds et al., In press; Spiering and Ashby, 2008). Depending on the test stimuli, these three decision bound models could make drastically different predictions, which would be of considerable theoretical interest.

4 | A NOTE ON EXCLUSION CRITERIA

Before presenting details of our next simulation, we briefly discuss a possible mis-apprehension that, in real experiments in the COVIS literature, participants using sub-optimum strategies would have been excluded from analysis through the application of a learning criterion.

In the category learning literature, participants who perform poorly (usually under a certain percentage) are often excluded (Newell et al., 2010). This is because researchers are typically interested in the process of learning, so if participants did not learn they are typically of little theoretical interest (Newell et al., 2010). However, effective and consistent exclusion of participants who are using sub-optimal strategies is unlikely to have occurred in the COVIS literature, for two reasons. First, it is relatively rare for studies in this literature to use a learning criterion (see Newell et al., 2010, for a discussion of the problems this can cause). Second, where learning criteria are used, they are typically too low to exclude plausible sub-optimal strategies. For example, using a unidimensional strategy in an information-integration category structure (such as in Fig. 1B) can result in participants scoring around 75%. This is higher than typical learning criteria in this literature. For example, Filoteo et al. (2010), Newell et al. (2010) and Smith et al. (2015) used learning criteria of 50%, 65% and 70% respectively. This problem is conceptually similar to one found in traditional analyses of the criterial-attribute category-learning task (Wills et al., 2015).

5 | SIMULATION 3: SMITH ET AL. (2014)

The previous two simulations demonstrated that DB analysis often misidentifies the strategies simulated participants use. Of particular concern for the COVIS literature is that rule-based and diagonal (GLC) strategies can be confused—these two types of strategy are often taken to index the action of qualitatively different learning systems (Explicit and

Procedural, respectively). If these strategy types were being misidentified in real experiments, it would undermine both the use of DB analysis and any experiment that relied on it.

However, the previous two simulations lacked one critical feature common to all real-world studies of category learning: accuracy scores. Typical categorisation studies focus on gross measures of performance (Kurtz, 2015), usually using strategy analyses only as a secondary measure, or manipulation check (e.g., Ashby and Valentin, 2016). Because sub-optimum strategies by definition cannot score as well as the optimum strategies, one might question whether it is possible to achieve the average performance scores seen in the literature while misidentifying participants' strategies. The current simulation demonstrates that it is.

Below, we use model-recovery techniques to demonstrate that current DB analyses misidentify participants' strategies in the context of levels of performance accuracy reported in published work. Further, we demonstrate that it is possible for all participants to be using rule-based strategies but yet to still find a) a behavioral dissociation (specifically, an interaction between an experimental manipulation and category structure), and b) that the majority of participants are (incorrectly) identified by DB analysis as using the optimum strategy for each category structure. By doing so, we hope to concretely demonstrate the risks inherent in the research methodology widely advocated in the COVIS literature and beyond (Ashby and Valentin, 2016; Smith et al., 2014, 2015).

The experiment we chose for this demonstration is by Smith et al. (2014); a recent, representative example of empirical work within the COVIS framework that has not been critiqued. This experiment investigated the effect of deferring feedback on category learning. Participants were randomly assigned to learn either a rule-based or information-integration category structure (as in Fig. 1) with one of two possible reinforcement schedules. In the immediate feedback condition, on each trial participants were shown a stimulus, then made their response and were immediately given corrective feedback for that trial. In the deferred feedback condition, the stimuli were shown to the participants in groups of six. The participants made responses for all six stimuli but only received corrective feedback at the end of the block. For instance, if the participant got half the categorization judgments correct, they would receive three "whoops" separated by 0.5s indicating correct responses followed by three buzzes separated by 4s timeouts indicating incorrect responses. Smith et al. found that learning of the rule-based category structure was unaffected by feedback timing, whereas learning of the information-integration category structure was "eliminated" (p. 454) with deferred feedback.

This experiment was chosen as a good test case for three reasons. First, it is representative of the majority of COVIS experiments (Ashby and Valentin, 2016) as it compares the effect of a manipulation on learning information-integration and unidimensional category structures. Second, the work reported in Smith et al. (2014) is interesting to simulate as it is representative of the direction that the role of DB analysis is beginning to take in newer COVIS experiments (see also, Smith et al., 2015). In these newer studies, the authors move away from using the DB analysis to ensure that participants were using the optimum strategy, and therefore category learning system, in each condition. Instead, they use the DB analysis to determine the strategies that participants use in order to discern whether deferring feedback alters the strategies participants use in "a theoretically meaningful way" (p. 452). Smith et al. (2014) concluded that deferred feedback pushed participants in the information-integration condition away from classification via the Procedural system toward classification via the Explicit system. These conclusions would of course be substantially undermined if their DB analysis failed to correctly identify the strategies participants used.

Thirdly, the possibility of a misidentification of participant strategies is theoretically interesting, as it opens the way for an alternative, single-system, account of their results. As previously discussed, verbal report data from Edmunds et al. (2015) indicate that participants sometimes learn information-integration category structures using complex, verbalizable rules—despite the DB analysis pointing toward procedural (GLC) strategies in these cases. Perhaps this is also happening in Smith et al. (2014)? Specifically, we hypothesize that the majority of participants in the immediate

information-integration category structure condition of Smith et al. are using a conjunction or another two-dimensional rule-based strategy, but this is mis-identified as an implicit (GLC) strategy by Smith et al.'s DB analysis. The possibility of this kind of mis-identification seems particularly acute in this study because those authors did not include a conjunction rule (or any other complex rule) in the set of models for their DB analysis. Research by Donkin et al. (2015) suggests that failing to include complex rules in a DB analysis increases the proportion of participants that are identified as procedural (GLC) responders.

5.1 | Method

To see whether it was possible that all the participants in Smith et al. (2014) were using rule-based strategies, we first generated a set of hypothetical participants. These participants' responses were generated from unidimensional and conjunction strategy models that best fit either the unidimensional or information-integration category structures used by Smith et al. We then added various levels of noise to these hypothetical participants and calculated their accuracy. Then we performed the DB analysis, which included three model types: unidimensional, diagonal (GLC) and random models. Note that although some simulated participants' responses were generated by a conjunction strategy, this strategy type was not included in the DB analysis. This was to keep the DB analysis as similar as possible to the one conducted by Smith et al. We then selected, using a process of trial and error, 21 simulated participants for each condition such that, as far as was possible, they had a) the same average accuracy as that reported in the experiment reported by Smith et al. (p. 451), b) the same number of "strong learners" (p. 451), and c) were identified by DB analysis as using the strategy types they reported (p. 452-453).

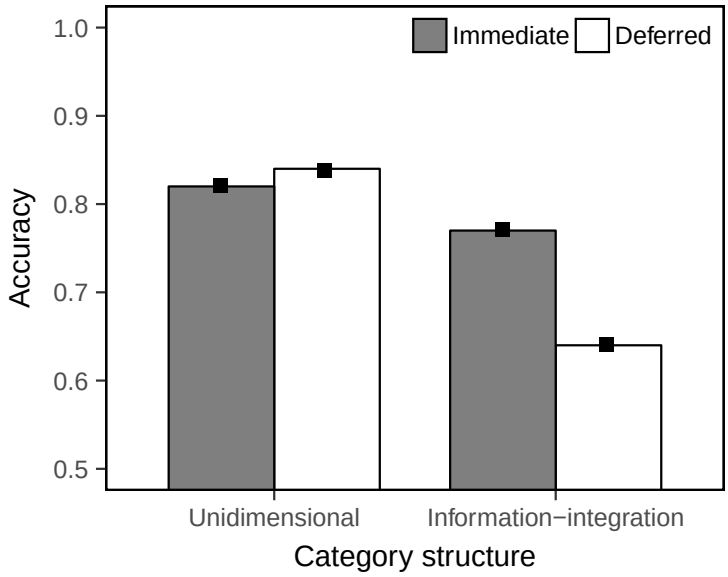

FIGURE 2 Simulation of Smith et al. (2014) using exclusively rule-based strategies. Bars represent the accuracy scores reported by Smith et al. (2014) and squares the accuracy of the simulated participants. Smith et al.'s report did not include an estimate of variability.

TABLE 3 The proportion of each type of strategy recovered for each type of generating strategy for a unidimensional category structure.

	Recovered strategies *(wBIC)*			
	UDx	**UDy**	**GLC**	**RND**
UD-Imm.	**13** (14)	**1** (1)	**1** (0)	**6** (6)
Gen. model: UD	14	0	0	4
Gen. model: CJ	0	1	0	2
UD-Def.	**15** (15)	**2** (2)	**0** (0)	**4** (4)
Gen. model: UD	15	0	0	4
Gen. model: CJ	0	2	0	0
II-Imm.	**0** (0)	**3** (3)	**16** (16)	**2** (2)
Gen. model: UD	0	0	0	2
Gen. model: CJ	0	3	16	0
II-Def.	**2** (2)	**13** (13)	**3** (3)	**3** (3)
Gen. model: UD	2	0	0	3
Gen. model: CJ	0	13	3	0

Strategies: UDx = Unidimensional based on the x-dimension, UDy = Unidimensional based on the y-dimension, GLC = General linear classifier (diagonal), RND = Random.

5.2 | Results

In addition to these hypothetical participants having the similar average accuracy (see Fig. 2), and distribution of DB-recovered strategies (see Table 3) as the real participants, it was also possible to replicate the statistical tests reported by Smith et al. with the simulated participants. For the simulated participants, the critical interaction between category structure and task was significant, $F(1, 80) = 30.54$, $p < .0001$. Furthermore, as in Smith et al. (2014), performance in the two rule-based conditions were statistically indistinguishable, $t(40) = 0.31$, $p = .760$, as was the comparison between the unidimensional and information-integration immediate conditions, $t(40) = 1.03$, $p = .310$, whereas, the difference between the two information-integration category structure conditions reached significance, $t(40) = 3.53$, $p = .001$

Table 3 shows that it is possible to generate a DB analysis closely similar to that reported in Smith et al. (2014), without assuming a dual-process model. Instead, all their participants could have been using rule-based strategies, with the participants apparently identified as using the Procedural System being the result of misidentification errors in their DB analysis. The one participant in Smith et al. (2014) that our simulation could not capture was in the unidimensional-immediate condition but was classified by Smith et al. (2014) as using a diagonal (implicit) strategy. This is an odd classification from the perspective of the COVIS model, as unidimensional classification should be explicit, according to this account. Our previous simulation found no examples of a simulated participant being recovered as GLC against the unidimensional category structure (see Simulation 2, Table 2). We hypothesize this recovery in Smith's data may have been due to a bias in that participant toward one category – a behavior not captured in our current recovery analyses. Clearly, there is the potential for further investigation here, but for now we would counsel against over-interpreting the

results of a single participant for whom the trial-level data has not been made available. Our overall conclusion from the current simulation is that we can produce accuracy and DB analysis results closely similar to those of Smith et al. (2014), without assuming a dual-process model.

6 | GENERAL DISCUSSION

6.1 | Summary

The influential COVIS model of category learning is supported by a great deal of behavioral evidence (Ashby and Maddox, 2005, 2011; Ashby and Valentin, 2016). Predominantly, this evidence comes from a single kind of experimental methodology, in which one examines the effect of some factor (e.g., concurrent load) on rule-based and information-integration category learning (Ashby and Valentin, 2016). COVIS predicts that its two learning systems (Explicit and Procedural) can implement different strategy types, and so each will learn one of these category structures better than the other. Critically, the validity of the inferences from this paradigm hang on correctly identifying the strategy each individual used to complete the learning task. This is because the experiments investigating COVIS do not directly control which system participants use to respond. Instead, they manipulate the category structures and hope that this encourages participants to use the optimum system, and thus the correct strategy, for that category structure. Of course, participants may continue to use a sub-optimum system for a particular category structure throughout the experiment. Thus, identifying the strategies participants use is crucial: if the participants are using the correct strategy for that category structure, then the experimenters assume that they must also be using the correct learning system for that structure (e.g., Ashby et al., 2002; Filoteo et al., 2010). Then, any differential effects of a manipulation on each category structure can be attributed to the existence of two systems of category learning, not differing numbers of sub-optimal responders.

In the COVIS literature, DB analysis is the key tool used to identify participants' strategies. However, as discussed in the Introduction, converging preliminary evidence casts some doubt on the adequacy of this use of DB analysis (Donkin et al., 2015; Edmunds et al., 2015). To investigate this possibility further, we conducted three model-recovery simulations. The first simulation used an information-integration category structure. In this simulation, we found around a third (37%) of simulated participants who used a conjunction (Explicit system) strategy were incorrectly recovered as using a diagonal (Procedural system) strategy. The second simulation used an unidimensional (rule-based) category structure. In that simulation, we found that nearly all (98%) of the simulated participants who used a diagonal (Procedural system) strategy were incorrectly recovered as using a unidimensional strategy (which is an Explicit system strategy).

In the third and final simulation, we provided a proof of concept that these sorts of misidentifications can result in a qualitative misinterpretation of experimental results. Specifically, we took a study used to support the idea of separate competing Explicit and Procedural category learning systems (Smith et al., 2014) and demonstrated that it was possible to closely model its means, inferential statistics and strategy analyses using just rule-based strategies. This raises the possibility that the participants in Smith et al.'s study were also only using rule-based strategies — an alternative, single-system, explicit account of their results. Furthermore, the results of this simulation demonstrate some of the pitfalls inherent in this experimental methodology, despite the fact that comparing rule-based and information-integration category structures has been strongly recommended by a number of researchers (Ashby and Valentin, 2016; Smith et al., 2014, 2015). It also highlights the need for researchers in this area to first demonstrate that their DB analysis accurately recovers its own models before using it to make theoretical claims.

6.2 | Implications for COVIS

Our three simulations demonstrate the existence of an inferential flaw in the experiments argued to support COVIS: the DB strategy analysis they employ is not accurate enough to act as a manipulation check. It cannot determine whether manipulating the category structure successfully elicited a corresponding switch in the categorization system underlying participants' responses. Consequently, it becomes difficult to judge whether a particular COVIS-supporting dissociation is due to the existence of two distinct learning systems, or rather due to participants using different explicit strategies to learn each category structure. This means that the conclusions of a large number of COVIS-supporting studies that rely on comparing rule-based and information-integration category structure have become uncertain (see Ashby and Valentin, 2016, for a list of such studies).

That being said, this is only one part of the evidence undermining the COVIS model. Although many studies have argued to support the COVIS model on the basis of dissociations between learning of rule-based and information-integration category structures, few of these remain un-critiqued. Briefly, the findings that dual-tasks interfere with rule-based but not information-integration tasks (Waldron and Ashby, 2001; Zeithamova and Maddox, 2006), have been shown to be due to including non-learners in the analysis (Newell et al., 2010). Kalish et al. (2017) and Lewandowsky et al. (2012) found that working memory demands do not dissociate between the two tasks, thus undermining other studies that claimed to find evidence that rule-based category learning relies on working memory but information-integration learning does not (e.g., Maddox et al., 2004c). Ashby et al. (2003) found that switching response keys interferes with information-integration but not unidimensional category learning. However, Nosofsky et al. (2005) demonstrated that this effect was more likely due to differences in cognitive complexity between the two category learning tasks. Similarly, Edmunds et al. (2015) found that Ashby et al.'s (2002) finding that feedback type affects information-integration but not rule-based learning could be explained by differences in the number of relevant stimulus dimensions between the tasks.

Other studies examining the role of feedback in COVIS have also been critiqued. For instance, the effect of reducing feedback time was found to dissociate between these two categorization tasks (Maddox et al., 2004a), but when the tasks were equated for perceptual difficulty the dissociation disappeared (Stanton and Nosofsky, 2007). Delaying feedback was also found to result in a dissociation (Maddox et al., 2003), but Dunn et al. (2012) showed that both using a non-Gabor pattern mask and presenting full-feedback removes the dissociation. Furthermore, Dunn et al.'s conclusions were strengthened by using state-trace analysis (Bamber, 1979; Loftus et al., 2004), which failed to find evidence of two independent systems (although see Ashby, 2014; Dunn et al., 2014; Yeates et al., 2015, for a discussion of the limitations of this analysis).

However, despite all these critiques, proponents of COVIS still publish studies looking to find a dissociation between learning of rule-based and information-integration category structures (e.g., Smith et al., 2014, 2015; Ashby and Vucovich, 2016) and strongly recommend these category structures as appropriate for studying category learning (Ashby and Valentin, In press). It is these most recent (and future) studies that the work presented here most targets.

6.3 | Implications for verbal report and DB analyses

Our simulations are consistent with the idea that participants can correctly report their categorization strategies for information-integration category structures. Edmunds et al. (2015, 2016) reported that participants learning information-integration category structures consistently reported using complex, rule-based strategies. In contrast, the DB analysis in those papers identified these participants as using a Procedural (diagonal) strategy. This posed something of a paradox, to which one resolution was that participants were inaccurate in their verbal reports. In the light of the current simulations another possibility is that participants were reporting accurately, but those using rule-based

strategies were mis-identified by DB analysis as using a Procedural (diagonal) strategy.

That being said, there is more work to be done before we can assume that the descriptions participants give best represent their performance in category learning tasks. For instance, it's possible that participants asked to describe their knowledge may be biased to produce rule-like responses, or that they are confabulating in some way. More pragmatically, the strategies that participants report can be somewhat vague. For instance, consider a participant who reported that they put"big stimuli in Category A, and small stimuli in Category B." This strategy is simple to classify as a unidimensional rule, but still does not provide details of what counts as "big" or "small." Furthermore, it is logically possible that strategies could exist on a continuum of "explicitness", with different learning conditions resulting in varying degrees of explicitness.

Therefore, in future work we hope to a) refine our procedure for eliciting verbal reports and b) take steps to evaluating their correspondence with the pattern of responding the participants give. Additionally, following Donkin et al. (2015), future work might explore using a greater variety of rule-based models within DB analysis that better represent the strategies that participants report using.

One potential issue with this approach is that complex rule-based models tend to have a lot of parameters. When combined with model-selection statistics such as AIC or BIC (Wagenmakers and Farrell, 2004), this may make it difficult for such models to be selected in a DB analysis. AIC and BIC, for any given level of fit to the data, favor the model with fewer parameters. This Occam's razor approach to model selection is popular but, as discussed by Wills and Pothos (2012), it really only makes sense in the context of fitting across the full range of known phenomena. If one choses to fit a subset of the data, for example to fit the binary categorization responses of the participant but to ignore their verbal reports, then the use of AIC/BIC exacerbates the risk that an overly-simple model is favored. We have no full solution to offer for this problem, but suggest (1) consideration of a more lenient model-selection statistic (e.g. maximum likelihood), and (2) that if models A and B have a similar likelihood for an individual's categorization responses, but model A is more consistent with the individual's verbal report than model B (perhaps as assessed by multiple independent raters), then model A should be favored.

In the context of our suggestion to use maximum likelihood, there is a certain irony in that fact that the most recent COVIS papers tend to use the Bayesian Information Criterion (e.g. Ashby and Vucovich, 2016; Spiering and Ashby, 2008) whilst earlier papers tended to use the Akaike Information Criterion (Ashby et al., 2002; Ell et al., 2009; Maddox et al., 2004b; Maddox and Ing, 2005). This is because both information criteria attempt to correct for model complexity using the number of parameters and the BIC penalizes high-parameter models more heavily than the AIC (Myung and Pitt, 1997). Of course, model complexity can be quantified in several different ways (Pitt et al., 2008), but using parameter number to distinguish among models here may have inadvertently increased the mis-classification of participants' strategies. Especially as the model with the highest number of parameters (the conjunction strategy) is mis-recovered most frequently. This hypothesis is supported by Donkin et al. (2015), who found that more people were identified as using rule-based strategies when model selection was done using AIC rather than BIC.

6.4 | Recommendations for future practice

How might the problems with the DB analysis of COVIS experiments identified in the current paper be reduced in future work? Here are our suggestions, which summarize a number of points we have made in the current article:

Include a conjunction strategy

Simulation 3 demonstrated that not including a conjunction strategy in the DB analysis can lead to qualitative misinterpretation of the data. Including a conjunction strategy may help; but note that DB analysis often fails to identify the use

of a conjunction strategy even when it theoretically could do so (Simulation 1), so this is not a complete solution.

Avoid single-dimension category structures

In Simulation 2, with a single-dimension category structure, DB analysis nearly always reported a single-dimension participant strategy, irrespective of the strategy the simulated participant actually used. A conjunction category structure may be a better choice for the "explicit" structure in a COVIS experiment, although model-recovery simulations should be performed to confirm this conjecture (see below).

Include a transfer phase

Most COVIS experiments test participants only on stimuli coming from the trained category structure. The results of Donkin et al. (2015) suggest that including a test phase where stimuli are drawn uniformly across the stimulus space may increase the accuracy of DB analysis, although again model-recovery simulations are needed to confirm this conjecture.

Collect and incorporate verbal report data

If the results of a DB analysis are at variance with what the participant says they were doing, do not immediately assume that the participant is wrong. Use your verbal report data to construct additional formal models for your DB analyses.

Perform model-recovery simulations

Having increased the set of candidate models, and improved the range of stimulus space covered by your test stimuli (see above), conduct model-recovery simulations to demonstrate the validity of your analysis method for your chosen models in the context of the experiment you intend to run.

Consider alternative methods of model selection

As discussed in the previous section, model selection based on AIC or BIC is not necessarily the most appropriate measure for DB analysis. Consider a range of model-selection statistics, such as maximum likelihood or hierarchical Bayesian methods that estimate posterior distributions of model probabilities and parameters (e.g., Kalish et al., 2017). If different statistics select different models, consider carefully why this might be — particularly if the models selected by AIC/BIC are inconsistent with other known data (e.g., verbal reports, reaction times, confidence ratings).

Publish your trial-level raw data

Not everyone will agree with the above recommendations. Either way, it is beyond dispute that DB analysis, unlike, for example, a Z-test, is not a single technique; it is a general approach whose specifics vary across time and between labs. In addition to the choice of model-selection statistic (see above), a number of other things differ between papers. For example, Ashby and Vucovich (2016) analysed only the last 100 training trials of their experiment, while Spiering and Ashby (2008) analysed each of four 150-trial blocks across their experiment. Casale et al. (2012) consider the general quadratic classifier to be an explicit rule-based strategy, while Ashby et al. (2001) consider it to be an implicit strategy. It is not possible to determine whether these differences between studies are important, because the trial-level raw data for these experiments have (at the time of writing) not been published. In future studies, publication of trial-level raw data would be highly desirable, as ideas of best practice seem likely to continue to change over time.

Use a wider range of (theoretically-motivated) models

One compelling but possibly incorrect intuition is that increasing the number of candidate models increases the difficulty of finding the generating model. It is, of course, the case that where all DB-analysis models provide an equally good fit to

the generating model, the probability of selecting the generating model reduces as the number of candidate models increases. On the other hand, if the generating model is not in the set, then the probability of selecting it is obviously zero. In practical terms, this second problem seems likely to be more severe than the first, particularly where the inclusion of candidate models can be justified on the basis of pre-existing theory or previous results. Nevertheless, if one is concerned about the dangers of including more models, then model-recovery simulations provide a good way of determining the validity of that concern for any given experiment and model set.

6.5 | Conclusions

In summary, the main take away messages are as follows. First, one should be very careful about drawing firm conclusions from the results of the DB analysis as currently put into practice. Second, the potentially low validity of this strategy analysis casts doubt on the COVIS research that relies on it as a manipulation check. Third, verbal reports are important. Fourth, and finally, there is more work to be done to determine the circumstances under which a decision-boundary modeling technique is reliable and valid.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Ashby, F. G. (2014) Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychonomic Bulletin & Review*, **21**, 935–946.

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

Ashby, F. G., Ell, S. W. and Waldron, E. M. (2003) Procedural learning in perceptual categorization. *Memory & Cognition*, **31**, 1114–1125.

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

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

— (2011) Human category learning 2.0. *Annals of the New York Academy of Sciences*, **1224**, 147–161.

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

Ashby, F. G., Paul, E. J. and Maddox, W. T. (2011) COVIS. In *Formal Approaches in Categorization* (eds. E. M. Pothos and A. J. Wills), chap. 4, 65–87. New York, NY: Cambridge University Press.

Ashby, F. G. and Soto, F. A. (2015) Multidimensional signal detection theory. In *Elementary Cognitive Mechanisms*, 13–34.

Ashby, F. G. and Valentin, V. V. (2016) Multiple systems of perceptual category learning: Theory and cognitive tests. In *Handbook of Categorization in Cognitive Science* (eds. H. Cohen and C. Lefebvre), 547–572. New York, NY: Elsevier, 2nd edn.

— (In press) The Categorization Experiment: Experimental Design and Data Analysis. In *Stevens' handbook of experimental psychology and cognitive neuroscience, Fourth Edition, Volume Five: Methodology*. New York: Wiley.

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

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

Bamber, D. (1979) State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, **19**, 137–181.

Carpenter, K. L., Wills, A. J., Benattayallah, A. and Milton, F. (2016) A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, **37**, 3557–3574.

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

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C. and Nosofsky, R. M. (2015) Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **41**, 933–948.

Dunn, J. C., Kalish, M. L. and Newell, B. R. (2014) State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby (2014). *Psychonomic Bulletin & Review*, **21**, 947–54.

Dunn, J. C., Newell, B. R. and Kalish, M. L. (2012) The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 840–859.

Edmunds, C. E. R., Milton, F. and Wills, A. J. (2015) Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, **68**, 1203–1222.

Edmunds, C. E. R., Wills, A. J. and Milton, F. (In press) Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology*.

Edmunds, C. E. R., Wills, A. J. and Milton, F. N. (2016) Memory for exemplars in category learning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (eds. A. Papfragou, D. Grodner, D. Mirman and J. C. Truesweell), 2243–2248. Austin, TX: Cognitive Science Society.

Ell, S. W., Ing, A. D. and Maddox, W. T. (2009) Criterial noise effects on rule-based category learning: The impact of delayed feedback. *Attention, Perception, & Psychophysics*, **71**, 481–489.

Estes, W. K. (1956) The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134–140.

Fific, M., Little, D. and Nosofsky, R. (2010) Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, **117**, 309–348.

Filoteo, J. V., Lauritzen, S. and Maddox, W. T. (2010) Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science*, **21**, 415–423.

Heathcote, A., Brown, S. D. and Wagenmakers, E.-J. (2014) *An Introduction to Good Practices in Cognitive Modeling*. New York, NY: Springer.

Helie, S., Turner, B., Crossley, M., Ell, S. and Ashby, F. (2017) Trial-by-trial identification of categorization strategy using iterative decision-bound modeling. *Behavior Research Methods*, **49**, 1146–1162.

Kalish, M. L., Newell, B. R. and Dunn, J. C. (2017) More is generally better: Higher working memory capacity does not impair perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **43**, 503–514.

Kurtz, K. J. (2015) Human category learning: Toward a broader explanatory account. In *Psychology of Learning and Motivation*, vol. 63, chap. 3, 77–114. Academic Press.

Lagnado, D. A., Newell, B. R., Kahan, S. and Shanks, D. R. (2006) Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, **135**, 162–183.

Lee, M. D. and Webb, M. R. (2005) Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605–621. URL: https://www.researchgate.net/profile/Michael{_}Lee/publication/7327314{_}Modeling{_}individual{_}differences{_}in{_}cognition/links/54fa9dd60cf20b0d2cb76680.pdf.

Lewandowsky, S., Yang, L.-X., Newell, B. R. and Kalish, M. L. (2012) Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 881–904.

Loftus, G. R., Oberg, M. A. and Dillon, A. M. (2004) Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, **111**, 835–863.

Maddox, W. T. (1999) On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354–74.

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

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

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

Maddox, W. T., Bohil, C. J. and Ing, A. D. (2004b) Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, **11**, 945–952.

Maddox, W. T., Filoteo, J. V., Hejl, K. D. and Ing, A. D. (2004c) 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**, 227–245.

Maddox, W. T. and 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.

Matsuki, K. (2014) GRT: General Recognition Theory. http://cran.r-project.org/package=grt.

Meeter, M., Myers, C. E., Shohamy, D., Hopkins, R. O. and Gluck, M. A. (2006) Strategies in probabilistic categorization: Results from a new way of analyzing performance. *Learning & Memory*, **13**, 230–239.

Myung, I. J. and Pitt, M. A. (1997) Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95.

Navarro, D., Griffiths, T., Steyvers, M. and Lee, M. (2006) Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, **50**, 101–122.

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

— (2011) Systems of category learning: Fact or fantasy? In *Psychology of Learning and Motivation*, vol. 54, chap. 6, 167–215.

Nosofsky, R. M., Stanton, R. D. and Zaki, S. R. (2005) Procedural interference in perceptual classification: implicit learning or cognitive complexity? *Memory & Cognition*, **33**, 1256–1271.

Nosofsky, R. M. and Zaki, S. R. (2002) Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 924–940.

Pitt, M. A., Myung, J. I., Montenegro, M. and Pooley, J. (2008) Measuring model flexibility with parameter space partitioning: an introduction and application example. *Cognitive Science*, **32**, 1285–1303.

Raijmakers, M. E. J., Dolan, C. V. and Molenaar, P. C. M. (2001) Finite mixture distribution models of simple discrimination learning. *Memory & Cognition*, **29**, 659–677.

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J. and Verfaellie, M. (2009) Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia*, **47**, 2995–3006.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Shanks, D. R. and St John, M. F. (1994) Characteristics of dissociable human learning-systems. *Behavioral and Brain Sciences*, **17**, 367–395.

Sidman, M. (1952) A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263–269.

Siegler, R. S. (1987) The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250–264.

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. a. and Ashby, F. G. (2014) Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, **25**, 447–57.

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G. and Church, B. A. (2015) The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, **77**, 2476–2490.

Spiering, B. J. and Ashby, F. G. (2008) Initial training with difficult items facilitates information-integration but not rule-based category learning. *Psychological Science*, **19**, 1169–1177.

Stanton, R. D. and Nosofsky, R. M. (2007) Feedback interference and dissociations of classification: evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, **35**, 1747–1758.

Wagenmakers, E.-J. and Farrell, S. (2004) AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**, 192–196.

Waldron, E. M. and 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.

Wills, A. J., Inkster, A. B. and Milton, F. (2015) Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, **80**, 1–33.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S. and Robinson, J. (2013) Is overall similarity classification less effortful than single-dimension classification? *The Quarterly Journal of Experimental Psychology*, **66**, 299–318.

Wills, A. J. and Pothos, E. M. (2012) On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, **138**, 102–125.

Yeates, F., Wills, A. J., Jones, F. W. and Mclaren, I. P. L. (2015) State-trace analysis: Dissociable processes in a connectionist network? *Cognitive Science*, **39**, 1047–1061.

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

APPENDIX

This appendix provides sufficient information to reproduce our model-recovery simulations from scratch. Alternatively, you may wish to consult our free and open-source implementation of these simulations, which is available at `www.willslab.org.uk/ply034`.

| Simulation 1

In Simulation 1, we looked to see how often the correct strategy type was selected from the DB analysis of responses generated from three common model classes: unidimensional (UD), conjunction (CJ) and diagonal (GLC).

For the *UD generator class*, responses were generated from a linear category boundary perpendicular to the x-axis, at $x = 50$, which is the optimal location of this boundary for all category structures in these simulations. Stimuli that fell to the left of the boundary were assigned 'Category A' responses and those that fell on the other side were given 'Category B' responses.

For the *GLC generator class*, responses were generated from a linear category boundary diagonally dividing the stimulus space. Then, the stimuli that fell above the boundary were given 'Category A' responses and those below 'Category B' responses. The diagonal boundary was fit to the category structure, so as to score the highest accuracy.

For the *CJ generator class*, responses were generated from two linear boundaries, one parallel and one perpendicular to the x-axis, that separated off a section of the stimulus space. Then, the stimuli that fell in that 'corner' of the space were given 'Category A' responses and the others 'Category B' responses. The position of the boundaries were fit to the category structure, so as to score the highest accuracy.

For each class of generating model, we generated a set of specific models by factorially varying the level of perceptual and decisional noise applied to the generating model. Perceptual and decisional noise are both central tenets of General Recognition Theory (Ashby and Gott, 1988; Ashby and Soto, 2015).

Perceptual noise reflects the fact that participants may not perfectly represent the stimulus at a perceptual level. We added perceptual noise to each stimulus sampled from a bivariate, normal distribution centred on the stimulus. Across specific generating models, the standard deviation of the perceptual noise took the values (0, 5, 10, 15, 20). The values of noise were chosen as approximate multiples of the standard deviation of the unidimensional category structure along the x-axis (4.04) and cover a range that might be reasonable. For the unidimensional and general-linear classifiers, stimulus coordinates were changed in a direction perpendicular to the decision boundary. For the conjunction strategy, the stimulus co-ordinates were changed in both the x- and y- directions with a covariance of 0.

Decisional noise reflects that fact that participants may not perfectly represent the location of their decision bound(s) in stimulus space. Decisional noise corresponds to variation in where the decision boundary passing through stimulus space lies. Decisional noise was modelled as a unidimensional normal distribution, orthogonal to the decision boundary. Across specific generating models, the standard deviation of the decisional noise took the values (0, 10, 20, 30, 40). These values were selected to cover a range from the absence of decisional noise, through to decisional noise that was somewhat larger than the distance between the two category structures (28.3). These levels of noise result in mean accuracy levels comparable to those found in published work for performance on an information-integration category structure. For each strategy type, mean accuracies were: $M_{UD} = 0.66$, $SD = 0.07$, $M_{GLC} = 0.75$, $SD = 0.12$, $M_{CJ} = 0.70$, $SD = 0.10$.

The final generated response for each stimulus was determined by seeing which side of the boundary (or boundaries in the case of the conjunction strategy) with decisional noise, the stimulus value with added perceptual noise was. For each strategy type and category structure, at each factorial combination of each level of perceptual and decisional

noise, 200 participants were generated. Each simulated participant classified the same 400 stimuli, sampled from the information-integration category structure provided by Smith et al. (2014).

Each of the simulated participants was then analysed using the DB strategy analysis procedures employed in the COVIS literature (Maddox and Ashby, 1993). The models used in the DB analysis were the unidimensional (UD), general linear classifier (GLC), conjunction (CJ) and random (RND) models.

The *UD strategy model* had two parameters: the value of the boundary and perceptual noise. Additionally, though only a unidimensional rule based on the x-axis was used in the generating models, both types of unidimensional rule were included in the model-fitting procedure. This was to account for the possibility in the information-integration category structure that either the unidimensional rule based on the x-axis or the unidimensional rule based on the y-axis might best fit the data.

The *GLC strategy model* had three parameters: the gradient and intercept of the boundary as well as perceptual noise.

The *CJ strategy model* had four parameters: the values of the two decision boundaries and two noise parameters, one for each stimulus dimension.

The *RND strategy model* assumed that, for each stimulus, category membership was assigned at random. The model has a single parameter: the probability of any stimulus being assigned to Category A.

The strategy that best represented each simulated participant's responding was defined as the one that minimized the Bayesian Information Criterion (Schwarz, 1978) for that participant. The resultant dataset from Simulation 1 (information-integration category structure) thus comprized 500 simulated participants per generating strategy (UD, CJ, GLC), who were each classified as using a UD, CJ, GLC or RND strategy by the DB analysis. Expressed as proportions, this is the data shown in Table 1. Model fitting was conducted with the aid of the *grt* package within the R environment (Matsuki, 2014). A further breakdown of the results in Table 1, by level of noise, can be found in the online materials

The high misrecovery rates for the CJ data-generating model may be due to competition from the simpler (i.e. lower-parameter) but non-data-generating CJ and UD models—the use of BIC as the model-selection criterion, standard in this field, penalises less simple models. The combination of this penalty with the poor discriminability of the generating strategies within the stimulus sets used leads to the simpler models often winning. As we covered in the General Discussion, including a transfer phase in future experiments may lessen this issue.

As a check of the robustness of our simulation, we re-ran it with another 200 simulated participants. The mean difference in the proportions reported in Table 1 between the two simulations was less than .01.

| Simulation 2

Simulation 2 was conducted in the same manner as Simulation 1, except that the unidimensional category structure from Smith et al. (2014) was used. The levels of noise result in mean accuracy levels comparable to those found in published work for each strategy type: $M_{UD} = 0.75$, $SD = 0.12$, $M_{GLC} = 0.75$, $SD = 0.12$, $M_{CJ} = 0.74$, $SD = 0.13$.

| Simulation 3

Simulation of the information-integration condition of Smith et al. (2014) involved selecting a subset of 21 simulated participants from the UD and CJ simulated participants of Simulation 1. The selection criteria were as described in the main text. Simulation of the unidimensional condition was performed in the same way, but using the simulated participants of Simulation 2. The CJ strategy model was not used in Simulation 3 for comparability with Smith et al. (2014), who did not use a CJ strategy model in their DB analysis of real participants.