Feature Inference in Perceptual Categorization

Emma Louise Morgan

A thesis submitted to the School of Psychology, Cardiff University, in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

September 2019

Supervisor: Mark Johansen

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed . . . . . (candidate) Date . . . . . .

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed . . . . . (candidate) Date . . . . . .

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed . . . . . (candidate) Date . . . . . .

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed . . . . . (candidate) Date . . . . . .

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed . . . . . (candidate) Date . . . . . .

Acknowledgements

First, I would like to thank the Cardiff University School of Psychology for funding this research and allowing me to ask the questions I wanted to know the answers to.

Second, I would like to thank my supervisor Mark Johansen who has taught me so much and who, through endless patience, has made me a better researcher.

I would also like to thank my friends at Cardiff University for lending me their ear when I needed it and giving me perspective.

Lastly, I would like to thank my family for their support, I couldn't have done this without them.

Summary

The ability to make inferences about the properties of a category instance based on knowledge of its category membership is a crucial cognitive ability. The purpose of this thesis was to evaluate feature inference learning of categories and feature inference decision-making for category instances in an attempt to clarify the nature of the category representation underlying feature inference. Specifically, three experiments evaluated feature inference learning of the classic Shepard, Hovland and Jenkins (1961) category structures compared to standard classification learning. The results supported a label bias hypothesis in terms of an advantage in feature inference learning tasks that allowed using label-based unidimensional rules over classification learning tasks that did not. However, both classification and feature inference learning resulted in predominantly rule representation, consistent with participants' self-reported learning strategies. A further five experiments evaluated feature inference decision-making in terms of analogues for standard categorical induction effects--notably premise typicality--in the perceptual categorization paradigm. However, there was no evidence of a premise typicality effect when similarity was controlled for. Possible conceptual and methodological reasons for failing to find this effect are discussed. While results do not support prototype representation as the underlying basis for feature inference, methodological explanations and/or a lack of power to detect a potentially small effect cannot be ruled out as explanations for the absence of premise typicality. The results of these eight experiments tentatively support a bias for label-based rules as the representation underlying feature inference learning and decision-making, but future research will need to more definitively differentiate this from prototype and exemplar representation.

Contents Page

Chapter One - General Introduction

1.1. Perspectives on Concepts

When interacting with complex environments, the ability to form categories is adaptively important because it facilitates the classifying of novel objects and events into categories and supports subsequent inferences of properties for those instances. Categories reduce time spent exploring the features of new objects, guide human behaviour in relation to these objects and aid communication (Murphy, 2002). For example, when encountering a new instance of the category 'apple', people do not have to wait for someone else to bite into it before inferring that it is edible. Classifying the novel object as an apple allows the generalization of the feature, 'edibility' from the known category of 'apple' despite never having seen that particular apple before. Making feature inferences about instances of categories is a crucial cognitive ability that pervades everyday interactions with the environment. An important perspective on categories is that they are fundamentally a way of organizing information to facilitate feature inference.

1.1.1 Rules and The Classical View

While it is clear that people have categories to represent sets of things in the world, assessing the nature of their mental representations is not straightforward, especially as people tend to report little direct awareness of how they categorize. What Smith and Medin (1981) called the Classical View of concepts is that each category is represented by a definition that contains all the necessary and sufficient features for an instance to be classified as a member of that category (Hull, 1920; Inhelder & Piaget, 1964; Smoke, 1932; etc.). However, this view has been shown to be inconsistent with many reported categorization findings (Hampton, 1979; Hampton, 1982; McCloskey & Glucksberg, 1978; Wittgenstein, 1953; etc.). For example, the Classical View assumes that all objects that meet the definition for a category are equally good members of that category and as such does not explain typicality effects. Typicality effects are

that some category members that have many features in common with many other category members are judged better examples of a category than instances with fewer features in common (Rosch & Mervis, 1975). For example, McCloskey and Glucksberg (1978) demonstrated that not all category members are good examples of a category. They tested participants on instances that were very typical of a category (e.g. chair in the furniture category), were intermediately typical (e.g. bookends) or were unrelated to the category (e.g. cucumber). They found that participants unambiguously categorized typical category members and clear non-members into the appropriate categories; however, participants were inconsistent in responding to the intermediately typical instances both between and within subjects across testing sessions. For example, when asked in separate sessions whether bookends were furniture, participants commonly gave different answers at different times. Some category members are better than others and some are not good examples of the category, suggesting that categories are not definitionally separate but are continuous.

The Classical View has been discarded as a theory of category representation because few, if any, real-world categories actually have necessary and sufficient conditions, i.e. sets of features that all instances have and all non-instances never have. Classically, Wittgenstein (1953) used the example of a definition for the concept of a game to highlight the difficulty of specifying necessary and sufficient conditions for most real-world categories. Instead, real-world categories tend to have family resemblances of features (Rosch & Mervis, 1975) where most instances of a category have many features in common, but not all, and few features in common with other categories, but not none. So, categories tend to have sets of features, each of which is semi-diagnostic of category membership. For example, the category ‘bird’ has many semi-diagnostic features in common across category members such as: having feathers, laying eggs, being able to fly, build nests, etc. So, feathers are strongly associated with the category ‘bird’ but removing the feathers from a bird still leaves a bird. Nevertheless, the

presence of feathers is a strong indicator that something is a bird. These are typical category features because they are shared by many category members, and birds that have few of these features are likely to be atypical category members. These differences in instance typicality are a key property of most real categories.

1.1.2. Exemplar and Prototype Representation

Accounting for the typicality structure of categories led to different theories of category representation, notably prototypes (Homa, Sterling, & Trepel, 1981; Smith, 2002; Smith & Minda, 2001; etc.) and exemplars (Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Nosofsky & Johansen, 2000; Nosofsky & Zaki, 1998; etc.). Exemplar representation is composed of individually remembered instances based on interactions with category members. So according to exemplar representation theory, when classifying a new instance as, for example, a dog, the mind determines the similarity of the new instance to previously stored instances of dogs. If the similarity is high, higher than to instances of other possible categories such as cats, then it will be classified as a dog. According to prototype representation theory, in contrast, categories are represented by an abstracted summary representation that combines the ideal/typical features of all category members rather than the instances themselves. That is, the category prototype is a kind of average or central tendency of the instances, a best instance. So, when classifying a new instance of the category ‘dog’, prototype theory says that the mind determines the similarity of the instance to various category prototypes and picks the one with the highest similarity. Both exemplar and prototype theory can account for category typicality effects, but they differ fundamentally in terms of what information the mind stores and uses to make new classifications: instances or abstractions.

Exemplar theory has been formalized in various mathematical models which use similarity to stored instances to categorize new instances (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; etc.). Medin and Schaffer (1978) proposed the Context Model. In this

model each feature value from each feature dimension is encoded as a configuration for every remembered exemplar in the category representation. To categorize a new instance, the model makes a comparison between the stored exemplars and the new instance on each feature dimension. The model sums across the differences for each feature dimension and uses an exponential decay function of that distance to specify an overall measure of similarity between the instance and a stored exemplar. This decay function means that similarity will be high when the instances are highly similar on all dimensions, i.e. the differences are small, but similarity will decrease quickly when even a small number of dimensional differences are large. The similarities for all instances are then summed to get an overall similarity to a given category's representation. The model does this for each possible category, and then the probability that a given instance is in a given category is its overall similarity to that category divided by its overall similarity to all relevant categories.

The Context Model (Medin & Schaffer, 1978) used a multiplicative rule that combined a measure of the differences between the instance and the representation and how important that difference is in terms of its impact on categorizing that instance as a member of a given category. This is in contrast to the common, previously used additive rule (Tversky, 1977), which calculates similarity as the sum of the dimensional features that match and mismatch. The key advantage of multiplicative similarity is that a single large dimensional difference between the category and the instance to be categorized can cause the model to say that the two are actually substantially dissimilar. This is important as some categories involve very typical features, the lack of which make an instance unlikely to be in the category despite potentially having many other features in common. In contrast, additive similarity in a model with the same large dimensional difference might only correspond to a slight decrease in the similarity between the category and the instance. Medin and Schaffer (1978) highlighted the importance of using multiplicative similarity through the example of the similarity between a mannequin

and a human being. If similarity was calculated by summing together all the features that mannequins and humans have in common, then similarity would be very high, and a mannequin might even be classified as a human. However, multiplicative similarity takes into account that an important feature in the similarity comparison is animation and because mannequins and humans do not share this feature, they are unlikely to be classified into the same category. Medin and Schaffer (1978) concluded that this exemplar model was more consistent with the data from their experiments than other models including a prototype model, thus supporting exemplar representation of categories.

The Context Model was further adapted by Nosofsky (1986) into the Generalized Context Model (GCM) which used Shepard (1957)'s multidimensional scaling-choice framework to explain similarity in the Context Model as inversely related to distances between instances in a constructed space; the bigger the distance between instances in terms of the larger differences between their feature values across stimulus dimensions, the lower the similarity. He also found an influence of selective attention to different feature dimensions within stimuli in terms of participants attempts to maximize accuracy in categorization tasks by paying greater attention to more diagnostic dimensions. Additionally, participants were adding exemplars that they experienced during the task to their exemplar-based representation. Overall, these models show that exemplar representations are capable of accounting for typicality effects (Busemeyer, Dewey, & Medin, 1984; Medin, Altom, Edelson, & Freko, 1982; Medin, Altom, & Murphy, 1984; Medin & Smith, 1981; etc.).

Prototype theory has also been formalized in various mathematical models which use similarity to an abstracted prototype to categorize new instances (Homa, 1984; Posner & Keele, 1968; Reed, 1972; Smith & Minda, 1998; etc.). Prototype models calculate similarity between test items and the category representation in much the same way as exemplar models but the category representations that instances are compared to are structured differently. Smith and

Minda (1998) used a prototype model with additive similarity to compare instances to the category prototype and applied attention weights to each of the feature dimensions within the category instances. They found that the prototype model fit the learning data better than an exemplar model early in the learning of a category with a large number of instances in it, though after initial learning, the exemplar model fit the data better. Overall, prototype models are, by their nature, compatible with typicality effects as the summary prototype representation is based on the typical features of a given category, so the more similar an instance is to the category prototype the more typical it is.

There has been a lot of research trying to clarify whether exemplars or prototypes are the basis for category representation (Ashby & Maddox, 1993; Malt, 1989; Minda & Smith, 2001; Palmeri & Nosofsky, 2001; Storms, De Boeck, & Ruts, 2000; Verbeemen, Vanpaemel, Pattyn, Storms, & Verguts, 2007; etc.) but there is little consensus on which is the correct representation. A variety of research has supported exemplar representation (Dopkins & Gleason, 1997; Nosofsky, Kruschke, & McKinley, 1992; Palmeri & Nosofsky, 2001; Shin & Nosofsky, 1992; Voorspoels, Vanpaemel, & Storms, 2008; etc.). For example, Voorspoels, Vanpaemel and Storms (2008) had participants rate the typicality of instances of 12 real-world categories such as birds. They fitted a prototype model and an exemplar model to individual and averaged participant data and found that the exemplar model fit both sets of data better than the prototype model. In contrast, there is a body of research claiming an advantage of prototype representation in explaining categorization performance (Minda & Smith, 2001; Minda & Smith, 2002; Smith, Osherson, Rips, & Keane, 1988; Smith, Redford, & Haas, 2008; etc.). For example, Minda and Smith (2002) reexamined the findings of 30 experiments that used the "5-4" category structure initially proposed by Medin and Schaffer (1978) and which provides support for exemplar representation. Minda and Smith found that when comparable prototype and exemplar models were used (where both models were based on a multiplicative

similarity calculation), the prototype model fit the data better than the exemplar model. However, Zaki, Nosofsky, Stanton, and Cohen (2003) argued that the specific version of the GCM that Minda and Smith (2002) used did not allow for differences in how deterministic responses were, as these are needed to account for individual participant data. Zaki et al. (2003) applied the full version of the GCM that allowed for differences in response determinism and found that the exemplar model provided a better fit of the data from Minda and Smith (2002) than the prototype model. Overall, exemplars and prototypes are very different perspectives on the underlying nature of category representation, but which is right has yet to be settled, due in part to a shift in the literature towards evaluating the multiple systems perspective.

Despite the fact that most categories do not have necessary and sufficient conditions, i.e. perfectly diagnostic rules, there are mathematical models of rule representation where category membership is at least partly determined by (possibly semi-diagnostic) rules (see Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Townsend, 1986; Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994; etc.). For example, Ashby et al. (1998) specified the COVIS model which is based on the assumption that there are two competing systems used in category learning: one explicit verbal system that works slowly, is consciously controlled and attempts to specify rules and the other which is implicit, similarity based and not consciously controlled. Even if this multisystem perspective turns out to be incorrect and/or not well differentiated from other kinds of representation, or worst of all hard to falsify, the data supporting the many models with some kind of rule representation certainly have attributes that implicate rules.

Rules, exemplars and prototypes are the dominant contenders for the basic representations formed when learning new categories. However, there are well-established influences on category formation that are harder to characterize in terms of these

representations. These influences include prior knowledge generally and more specifically, the causal relations contained within that knowledge.

1.2. Prior Knowledge Influences in Categorization

The effect of prior knowledge in categories and on category learning is pervasive (Gratton, Evans, & Federmeier, 2009; Kaplan & Murphy, 2000; Lewandowsky, Kalish, & Griffiths, 2000; Palmeri & Blalock, 2000; Rips, 1989; Spalding & Murphy, 1996; Wattenmaker, Dewey, Murphy, & Medin, 1986; Wisniewski & Medin, 1994; Ziori & Dienes, 2008; etc.). For example, Rips (1989) used a task in which an object was described to participants as being between the size of a quarter (U.S. 25 cents) and a pizza, and participants rated whether the object was more likely to be a quarter or a pizza. Participants preferentially responded that the object was a pizza even though it was more similar in size to quarter, based on their prior knowledge that pizzas vary in size, but quarters do not. This demonstrates an effect of prior knowledge on the categorization of new instances.

Prior knowledge can improve category learning performance in various ways. For example, Pazzani (1991) evaluated a common finding (Conant & Trabasso, 1964; Haygood & Bourne, 1965; Hunt & Hovland, 1960; etc.) that conjunctive categories can be learned faster than disjunctive categories. Conjunctive categories are based on two features that must occur together e.g. in a category of balloons a given balloon instance must be small *and* yellow. Disjunctive categories are based on two features, one of which must occur or both for classification into a given category e.g. that a balloon must be either small *or* yellow *or* both. Pazzani asked one set of participants to predict whether an image of a person doing an action to a balloon was an alpha or not (a blank categorization word) and asked another set of participants to predict whether the balloon in the image would be blown up or not. The conjunctive concept to be learned was size of balloon = small and colour of balloon = yellow and the disjunctive category to be learned was age = adult or action = stretching a balloon.

With the expectation and prior knowledge given by the instruction that the goal was to blow up the balloon, participants were able to learn the disjunctive category easier than the conjunctive. This suggests that the inclusion of prior knowledge in a task can reorder classic categorization learning findings.

Prior knowledge also facilitates the attention to and usage of features that are marked by that knowledge to be important. For example, Lin and Murphy (1997) had participants learn novel categories and then gave them background knowledge about the importance of each feature in the category. If, at test, novel instances did not have the features that their knowledge marked as important, then the participants were less likely to categorize that instance as a member of that category. In contrast, Heit (1998) found that when participants were given a list that described a person, features that were incongruent with prior knowledge had a greater impact on transfer performance than congruent features. Heit argued that participants tried to explain how these features fit with the rest of the instance and therefore attended to these features more. This suggests that prior knowledge can impact what information is attended to in a categorization task, both in terms of more attention to features consistent with that knowledge but also, sometimes, more attention to features inconsistent with that knowledge.

Murphy and Allopenna (1994) evaluated the nature of the features associated with prior knowledge and found that categories of meaningful phrases were as hard to learn as categories of meaningless symbols, suggesting that it is not the meaningfulness of features per se that gives prior knowledge its advantage. Rather, they found that the relationships between the features impacted learning, that is, features connected by a theme were easier to learn. For example, the features: green, made in Africa, lightly insulated and can drive through jungles can all be connected by the theme ‘jungle vehicle’ which aids understanding of the category and how the features relate to it. This suggests that prior knowledge promotes easier category learning by connecting category features together.

Additionally, there is an impact of *when* prior knowledge is introduced on the ease of learning a category. For example, Wisniewski (1995) found that when participants were given knowledge about the category before learning, performance was significantly better than if the knowledge was given after learning but before testing. Further, Heit and Bott (2000) found that the learning of features that were critical to the categorization of instances was facilitated by prior knowledge whilst the learning of non-critical features was not. For some categories of stimuli, this advantage was present from the first learning block, suggesting that the impact of prior knowledge occurs very early in category learning.

Categorical induction is a particular kind of knowledge effect assessed in terms of feature inferences about instances known to be in a particular category, and knowledge of category membership tends to have a strong influence on feature inference. For example, Kalish and Gelman (1992) showed an effect of prior knowledge in categorical induction tasks. They tested children on inductions for features of items with two relevant categories such as 'wooden pillows' and found that the participants could make inductions of properties based on the relevant category. For example, if making the induction about whether the wooden pillow would be hard or soft, children were able to infer that it would be hard based on the 'wooden' feature despite the fact that pillows are typically soft. Further, Ross and Murphy (1999) investigated whether using certain types of categories in an inference task promoted certain kinds of inferences. They presented participants with a target instance (e.g. cereal) and then two comparisons: one that was in the same taxonomic category (e.g. noodles, as both are in the category 'grains') and one that would be used in the same situation as the target (e.g. milk). They then asked two questions: which food was likely to share an enzyme with the target (biochemical comparison) and which was likely to be eaten at the same meal (e.g. situational comparison). They found that participants rated the same category option as more likely when making a biochemical comparison but rated the same situation option as more likely when

making a situational comparison. So prior knowledge about categories and situations influences which are used when.

Not surprisingly, prior knowledge effects particularly occur in subject area experts (Johnson & Mervis, 1997; Medin et al., 2006; Shafto & Coley, 2003, Vitkin, Coley, & Hu, 2005; etc.). For example, Proffitt, Coley, and Medin (2000) gave induction problems to tree experts relating to the susceptibility of certain trees to given diseases, and they found that the experts responded based on reasoning about how those diseases could be transmitted, the thickness of the specified tree's bark, etc. rather than basing their reasoning on the typicality or diversity of the given instances. This suggests that with well known, real-world categories, specific prior knowledge tends to influence categorization more than just similarity, typicality or diversity.

The influence of knowledge on induction can be included in categorization models (Heit & Bott, 2000; Mooney, 1993; Rehder & Murphy, 2003; etc.). For example, Heit and Bott (2000) proposed a connectionist model in which the hidden layer, between the feature-based input layer and the category response output layer, contained the prior knowledge hard coded. Learning with this model can occur directly as an association between the features (input) and the category response options (output) or indirectly using the prior knowledge in the hidden layer. Another example of a model of prior knowledge is the Induction Over the Unexplained (IOU) model that Mooney (1993) proposed. The IOU model is an artificial intelligence learning model which categorizes by first using all the features of a concept that match prior knowledge. When all the prior knowledge has been applied and if there are still features/instances left then the remaining information is passed on to an empirical system that will try to find similarities to add to the overall concept. But one of the challenges with such models is that the knowledge they contain needs to be accurate, and knowledge in many domains can be quite complex.

Causality is a specific form of prior knowledge i.e. how one feature might causally be connected to another. Causality can influence categorization performance (Ahn, Kim, Lassaline, & Dennis, 2000; Lassaline, 1996; Lien & Cheng, 2000; McNorgan, Kotack, Meehan, & McRae, 2007; Rehder & Burnett, 2005; Rehder & Hastie, 2004; Rehder & Kim, 2006; Rottman, Gentner, & Goldwater, 2012; etc.) and causal knowledge of a category can impact inductions of new features (Rehder & Hastie, 2001). For example, Rehder and Hastie (2004) found that the existence of causal relationships between category instances strengthened the coherence of that category which strengthened inductions based on those instances. Participants were presented with a series of category instances such as 'Lake Victoria Shrimp', were told that a given instance had a novel property and were asked to make an induction as to what proportion of the rest of the category was likely to have that feature. For a Lake Victoria Shrimp, two of the four features participants were told it possessed were, 'Has high amounts of ACh neurotransmitter' and 'Has a long-lasting flight response'. Some participants were additionally told of a causal link that 'A high quantity of ACh neurotransmitter causes a long-lasting flight response'. Participants who were given this additional information rated exemplars that were consistent with the causal relations as more likely to generalize to the rest of the category than participants who did not receive such information. This shows that causality can impact inductive strength, potentially above and beyond other factors such as typicality, but as with other kinds of knowledge, causal interrelationships can be complex.

The focus of Experiments 4-8 in the present research was on assessing feature inference as a kind of categorical induction, that is, on attribute inference in the context of categories as a particular kind of knowledge effect. However, there are two key challenges for characterizing knowledge influences on category learning and decision-making: one is that there is still relatively little consensus on the basic nature of category representations in terms of when, where and how people use rules versus prototypes versus exemplars. The second is that the full

extent of knowledge effects on real-world categories is difficult to characterize accurately because of their complexity. So, the present research took a simplified approach to these difficulties by using the perceptual categorization paradigm.

1.3. Perceptual Categorization

1.3.1. The Paradigm

Many of the categories people learn are based on visual exposure to category instances with attached conceptual labels, e.g. cat, tree, cloud, etc. and as such the perceptual categorization paradigm, which uses novel, carefully controlled stimuli and newly constructed categories (for examples see Griffiths, Hayes, & Newell, 2012; Honke, Conaway, & Kurtz, 2016; Johansen & Kruschke, 2005; Love, 2002; Nosofsky & Zaki, 2002; Yamauchi & Markman, 1998; Zeigler & Vigo, 2018) is a way to assess the basic mechanisms of category learning and feature inference. So, in contrast to real-world categories, which are complex and interconnected with background knowledge, simple perceptual categories can be constructed to test how people learn and represent new concepts and make inferences about features while limiting the complex interplay with background knowledge because the categories are new.

1.3.2. Classification and Feature Inference Representations and the Impact of Category Labels

There are at least two ways to learn about categories via feedback: one is by classification learning, assigning an instance to a category and then being told the correct category, e.g. classifying a small, furry animal as a cat, not a dog. Another is by feature inference learning, inferring features of known category instances and being told the correct feature, e.g. inferring a cat is likely to purr if you pet it (rather than bite). A key difference between these two learning tasks is the presence of category membership information as essentially part of the stimulus in feature inference. In perceptual category learning, this category information commonly takes the form of a verbal category label. This difference in available information suggests the possibility that classification and feature inference learning

and decision-making result in fundamentally different category representations because of the presence of the label in feature inference.

1.3.2.1. Classification and Feature Inference Learning

The presence of verbal category labels impacts categorization performance. Yamauchi and Yu (2008) found that when verbal labels indicated the category membership of an instance rather than indicating the description of an additional feature, participants were more likely to make feature inferences consistent with the labels. Further, Yamauchi, Kohn, and Yu (2007) found that verbal category labels that conveyed category membership information were viewed earlier in a trial and more often than feature based labels that did not convey category membership information. Finally, Lupyan, Rakison, and McClelland (2007) showed that having labels present during a task aided categorization even if the labels were redundant with other information, compared to not including a label in the task at all.

Category labels are closely tied to the functionality of categories. Yu, Yamauchi, and Schumacher (2008) argued that labels have an impact on categorization tasks by highlighting the interrelatedness of features and subsequently increasing the perceived similarity across features within a category. Johansen, Savage, Fouquet, and Shanks (2015) concluded that the dominant influence of the label on feature inference was due to it being more salient than other features. Taken together, these studies suggest that the label is distinct from all other features presented during a categorization task and produces differences in learning and responding from tasks in which the label is not part of the stimulus such as in classification tasks.

The influence of category labels on feature inference suggests category representations based on them. Consistent with this, Yamauchi and Markman (1998) hypothesized that feature inference learning tasks, which include the category label, encourage learning the internal structure of each category and the typicality of individual features within a category, which induces prototype representation. Classification learning, in contrast, encourages learning the

differences between categories. Anderson, Ross, and Chin-Parker (2002) further supported Yamauchi and Markman's hypothesis: participants who had completed a feature inference learning task performed better on single feature classifications than full instance classifications. This supports the hypothesis that feature inference encourages the learning of prototypical features and thus encourages prototype representation. Johansen and Kruschke (2005) also contrasted these kinds of learning on the 5-4 category structure from Medin and Schaffer (1978) and argued that feature inference encouraged the formation of a set of label-based rules that sometimes mimicked prototypes and sometimes didn't. From this they suggested that feature inference learning does not induce prototype representation per se but rather a tendency to form rule representations based on the category labels, in contrast to classification learning.

Similarly, Sweller and Hayes (2010) argued that there is a difference in the content of the representations formed through classification and typical feature inference learning but not between classification and mixed feature inference learning. They distinguished between these two types of feature inference learning as: typical feature inference where learning occurs through querying only the typical features of the category instances, and mixed feature inference where learning occurs through the querying of both typical and atypical features. They argued that typical feature inference produced a different representation from classification but as a result of the methodological artefact of only asking for typical feature inferences. In contrast, they argued that there is no representational difference between classification and mixed feature inference learning as both promote the incorporation of both typical and atypical features and both lead to exemplar representation. However, Chin-Parker (2011) used a feature inference task that queried both typical and atypical instances and found that participants struggled to learn the structure to over 75% accuracy in the last learning block. He took this as evidence that participants were not storing exemplars, which should potentially allow near perfect performance. However, if they were learning a prototype, the prototype

inconsistent features would remain hard to learn. So, the poor learning was taken as evidence of a bias for prototype representation.

There is little consensus in the field about the representation underlying feature inference learning. As previously stated, Sweller and Hayes (2010) argued for exemplar representation for feature inference learning, Yamauchi and Markman (1998) made arguments consistent with prototype representation, and Johansen and Kruschke (2005) concluded that the representation is based on rules that use the category label. The variety of proposed representations for feature inference learning, while not providing a consensus on the true nature of the representation, have nonetheless emphasized the importance of evaluating the basis for feature inference.

1.3.2.2. Classification and Feature Inference Decision-Making

Similar to classification and feature inference learning, classification and feature inference decision-making have been argued to be different, with feature inference especially influenced by category membership information. Gelman and Markman (1986) showed that feature inference decision-making for real-world categories was more heavily influenced by category membership than by perceptual similarity: they showed four-year-old children two pictures of category instances and described an associated property. They also showed the children a third instance that had the same category label as one instance but was perceptually more similar to another and asked which associated property the queried instance was likely to have. For example, a child might have been shown a tropical fish and been told that it could breathe underwater, been shown a dolphin and been told it jumps up out of the water to breathe and then been shown a shark as the testing item. The shark had the same category label as the tropical fish, ('fish') but was perceptually more similar to the dolphin. The children were asked if the shark could breathe underwater or if it has to jump out of the water to breathe. They found that children preferentially used category membership information to make the feature

inference inductions even when perceptual similarity indicated a different response. Similarly, for adults, Yamauchi and Markman (2000) found that feature inferences were more likely to be determined by a category label than by perceptual similarity. From this they argued for the special status of the category labels. Taken together, prior research suggests that the presence of the label in feature inference plausibly induces a difference in focus compared to classification. The present research assessed feature inference and the potential impact of the presence of the category label in the classic category structures from Shepard et al. (1961) as well as in a variant of the family resemblance structure.

1.4. Shepard, Hovland and Jenkins (1961)

The category structures specified by Shepard et al. (1961) are a benchmark assessment of category learning in the perceptual categorization paradigm. The reason these category structures are important is that they represent an evaluation of the learnability of what are among the simplest, non-trivial categories. Shepard et al. (1961) evaluated the relative learnability of all possible category structures formed with eight instances, equally split into two categories, with instances composed of features from three binary-valued dimensions. There are six basic category structure types that are consistent with these constraints, Figure 1. In the figure, each type is a cube with the specific instances of the A and B categories at the corners of the cube and the edges indicating the three feature dimensions and thus the features composing each instance. The Type I structure can be learned using a rule on a single dimension, dimension one, that allows one feature to be exclusively associated with one category and the other feature to be exclusively associated with the other category e.g. the feature ‘square’-shaped only occurs in instances of category A, and the feature ‘triangle’-shaped in category B. The Type II structure, Exclusive-Or, can be learned using a rule based on the configuration of the first two dimensions e.g. instances with features ‘white’ and ‘square’ or ‘black’ and ‘triangle’ are occurrences of category A while ‘black’ and ‘square’ or

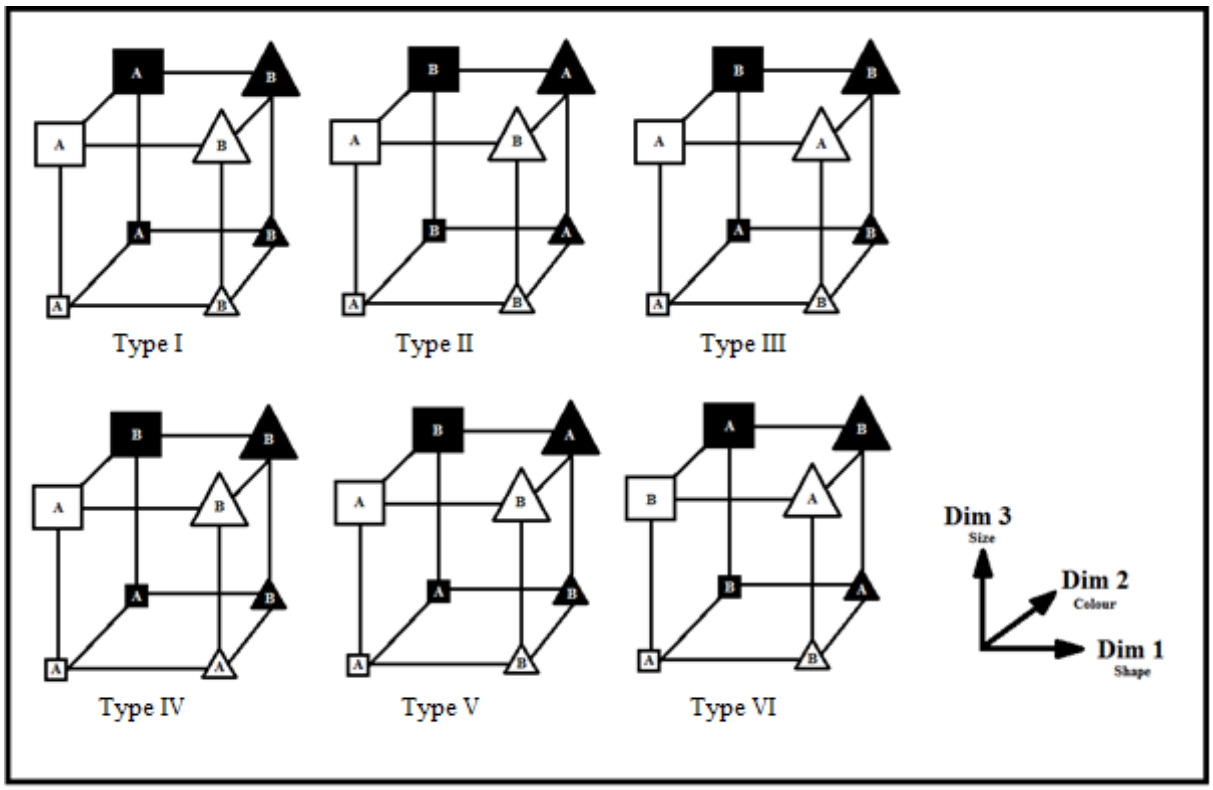

Figure 1. The six types of category structures from Shepard et al. (1961). The diagram in the bottom right shows the assignment of the three stimulus dimensions to each dimension of the abstract category structures i.e. the cubes. Each corner of a cube represents a category instance as composed of a feature value from each of the three binary-valued stimulus dimensions. 'A' labels indicate instances of one category and 'B' labels indicate instances of the other category. (Adapted in part from Kruschke, 1992 and Shepard et al., 1961).

'white' and 'triangle' are occurrences of category B. For Types III, IV and V, learning a rule on the first dimension allows correct categorization of six out of the eight instances, but the remaining two exceptions have to be handled in some other way e.g. for type V, category A instances are either 'square' or 'large' 'black' and 'triangle' and category B instances are either 'triangle' or 'large' 'black' and 'square'. Finally, for Type VI, each category instance can be memorized, or the structure learned in terms of the Odd-Even rule. The Odd-Even rule requires memorization of a single instance and if another instance varies from that instance by one feature or all three features then the correct category is the opposite of the category for the

memorized instance. If the new instance varies by two features, then the correct category is the same as the memorized instance (Shepard et al., 1961).

The classic findings of Shepard et al. (1961) were systematic differences in the learning difficulty for these six category structures. The Type I category structure was the easiest to learn, Type II was more difficult, Types III, IV and V were equally difficult but all harder than Type II, and Type VI was the most difficult; in summary I<II<III=IV=V<VI. This pattern of learning has been replicated and evaluated many times (Edmunds & Wills, 2016; Griffiths, Christian, & Kalish, 2008; Kruschke, 1992; Kurtz, 2007; Love, Medin, & Gureckis, 2004; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Nosofsky, Palmeri, & McKinley, 1994; Rehder & Hoffman, 2005; Smith, Minda, & Washburn, 2004; Žauhar, Bajšanski, & Domijan, 2016; etc.) though some studies have not clearly differentiated the learning between some specific types (Kurtz, Levering, Stanton, Romero, & Morris 2013; Lewandowsky, 2011; Love, 2002; Žauhar, Bajšanski, & Domijan, 2014; etc.). The key conclusion of Shepard et al. (1961) was that this pattern of learning difficulty reflects the complexity of the rules that allow accurate performance, more complex rules are harder to learn, and as such these results clarify the cognitive mechanisms involved with basic category learning.

Nosofsky et al. (1994) replicated Shepard et al. (1961) with similar stimuli and found the same ordering of the types, and this represents a kind of canonical replication. However, deviations from the standard type ordering have been found as a result of initial task instructions and the specific stimuli used: Nosofsky and Palmeri (1996) compared integral dimension stimuli to the classic separable dimension stimuli used by Shepard et al. (1961) and found that Type II was more difficult than Types III and IV and not significantly different from Type V. Love (2002) evaluated these types using incidental unsupervised learning and found that Type IV was easier than Type II. Kurtz et al. (2013) evaluated a variety of manipulations and found that the relationship of Type II to the other types can be changed in various ways

and questioned the universality of the ordering of Type II in the classic results. Taken together, these results suggest that although there are various influences on the exact ordering and differentiation of the intermediate types, the overarching pattern of Type I being easiest and Type VI being hardest and the other types being intermediate is fairly reliable.

Whilst this pattern of learning is reasonably well established for classification learning there have been to our knowledge no attempts to assess feature inference in terms of the learning and subsequent representation underlying it for these structures. The results of learning the Shepard et al. (1961) structures by feature inference, are reported in Chapter 2 while the experiments in Chapters 3 and 4 used a variant of the family resemblance category structure to assess categorical induction as feature inference decision-making.

1.5. Categorical Induction

Categorical induction involves making judgements about unknown features of a category instance based on known features of known category members from previously known categories (Gelman & Markman, 1986; Heit, 2000; Medin, Lynch, Coley, & Atran, 1997; Proffitt et al., 2000; Rips, 1975; Rips, 2001; etc.). An example of a categorical induction argument taken from Hayes, Heit and Swendsen (2010) is, 'Sparrows have property X Therefore Geese have property X'. This argument includes already known categories of birds, however implies a generalization of an unknown feature (property X) from one known category member to the other (sparrows to geese). The common response measurement for these arguments is a rating of the likelihood of the conclusion being true (that geese have property X) given that the premise is true (that sparrows have property X). These likelihood ratings can be used to measure the strength of an inference for an unknown attribute.

1.5.1. Comparison of Categorical Induction and Feature Inference

Attribute inference in categorical induction and feature inference in perceptual categorization are similar. Both processes are based on using category knowledge to make

inferences about what feature an instance might have, but the origin of the knowledge is usually different. Categorical induction normally uses known categories such as birds or mammals, that is, rich and complex real-world categories acquired over a lifetime (see Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975; Smith, Shafir, & Osherson, 1993; etc.). Feature inference typically uses newly learned, constructed categories (see Griffiths et al., 2012; Johansen & Kruschke, 2005; Murphy & Ross, 1994; Yamauchi, Love, & Markman, 2002; etc.). As described earlier, this learning about category instances can be done by classification where an instance is presented and participants classify it into one of several categories and then receive feedback, or by feature inference in which participants chose between possible features for a known category instance and receive feedback. Testing subsequently occurs via classification or feature inference without feedback.

Both categorical induction and feature inference ask participants to make a response about an attribute/feature that is hidden or not visible, though the nature of these responses is different, a rating of argument strength in categorical induction versus a chosen feature in feature inference. Nevertheless, these two kinds of responding should be related: if a participant believes that one argument is stronger than the other as manifested through a difference in ratings on the likelihood scales, the participant would plausibly choose the response/feature associated with the stronger argument when faced with binary responses options that are essentially pitting the two arguments against each other in a forced choice. Overall, the strong similarities between these two paradigms suggest that effects found in the categorical induction paradigm should also occur in the more methodologically controlled perceptual categorization paradigm.

The categorical induction paradigm has a corpus of well-established empirical effects, including premise typicality, premise conclusion similarity, etc. as described in detail below. However, the category representations underlying these effects are unclear, in part because the

categories themselves, e.g. birds, are so complex. Establishing these effects in the perceptual categorization paradigm would allow an assessment of the representations underlying these effects using the well specified representation models described above, i.e. prototype and exemplar models.

1.5.2. Categorical Induction Effects

1.5.2.1. Premise Typicality and Typicality

Premise typicality is one well-established effect in the categorical induction paradigm; the more typical a premise item is, the stronger the conclusion drawn from it for other instances, i.e. the greater the judged argument strength (Rips, 1975). An example from Hayes et al. (2010) is, “Sparrows have property X Therefore Geese have property X" is judged to be a stronger argument than, "Penguins have property X Therefore Geese have property X.” The first argument is judged to be stronger because a sparrow is a more typical exemplar of the bird category than a penguin and shares more features with other category members. So, an argument based on a premise using a typical category member is judged to be stronger than an argument based on an atypical category member.

Category typicality effects are common and well documented (Light, Kayra-Stuart, & Hollander, 1979; Lin, Schwanenflugel, & Wisenbaker, 1990; McCloskey & Glucksberg, 1978; Medin & Schaffer, 1978; Nosofsky, 1988; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976; Rothbart & Lewis, 1988; Spalding & Murphy, 1999; etc.). As discussed above, Rosch and Mervis (1975) specified typicality in terms of features shared across category instances: a category instance is most typical when it has many features in common with other members of the same category and few features in common with members of other categories. One explanation for the premise typicality effect then is that the feature queried in an argument can be better generalized across a category by an instance that has many features in common with other category members than by an instance with fewer features in common.

1.5.2.2. Other Categorical Induction Effects

There are many other well documented effects in categorical induction distinct from typicality and premise typicality. One such effect is premise conclusion similarity where the more similar the conclusion is to the premise, the stronger the argument is judged to be. For example, 'Leopards have property X Therefore Lions have property X,' is judged to be a stronger argument than, 'Leopards have property X Therefore Koalas have property X,' because leopards are more similar to lions than to koalas (Hayes et al., 2010). This is different from tests of premise typicality which attempt to hold similarity constant across the premises such that only the typicality of the instances' influences responding.

Premise diversity is another common effect in categorical induction where the greater the extent to which the premises of an argument "cover" their parent category, the stronger the conclusion is judged to be. An example of a diverse argument taken from Hayes et al. (2010) is, 'Lions and Mice have property X Therefore Mammals have property X.' This is judged to be a stronger argument than the less diverse, 'Lions and Tigers have property X Therefore Mammals have property X' because lions and mice cover the category 'mammals' better than lions and tigers, which are both cats.

The inclusion fallacy is another reasonably well-established categorical induction effect that is based on the idea that making a conclusion instance more general can strengthen an inductive argument. It occurs when a conclusion that covers the whole of a category is judged as stronger than a conclusion that is a specific member of that category. For example, 'Crows have property X Therefore Birds have property X,' is judged to be a stronger argument than, 'Crows have property X Therefore Ostriches have property X,' (Hayes et al., 2010). The category 'birds' includes ostriches as well as many other bird instances and therefore the argument based on the bird conclusion requires far more birds to have property X (weakening the argument as it's likelihood decreases) than simply requiring a single bird instance (ostrich)

to also share the feature. So, choosing the bird conclusion as the stronger argument demonstrates a fallacy in that the general conclusion is judged as a stronger argument than the specific conclusion despite its implied reduced likelihood. Shafir, Smith and Osherson (1990) argued that this fallacy is due to typicality. So, for the current example, crows are more typical of the category bird and less typical in relation to the category ostrich, therefore the strength of the argument is based on the strength of the typicality relations between the premise and the conclusion. As the premise and general conclusion relation has stronger typicality, this is rated as a stronger argument despite the specific conclusion being logically more likely.

1.6. Overview of the Thesis

1.6.1. Thesis Motivation

The prominent position of typicality in classification, feature inference and categorical induction tasks naturally invokes prototype representation. A summary from Murphy (2002, p. 265) emphasizes this, "If read literally, almost all the work on category-based induction takes a prototype view of concepts. This is not to say that researchers on induction propose a specific category representation along with a learning rule. However, the talk about concepts is almost inevitably one in which a concept has a summary representation." So, part of the original motivation for the present research was to be able to discriminate prototype and exemplar representations using feature inference tasks.

1.6.2. Experiment Summary

The purpose of this research was to evaluate feature inference in perceptual categorization as a mode of learning about categories and as a mode of decision-making using categories to clarify the underlying category representations. The first three experiments (Chapter 2) evaluated and clarified feature inference learning of the classic Shepard et al. (1961) category structures in contrast to classification learning. The second three experiments (Chapter 3) assessed analogues of premise typicality as feature inference in perceptual

categories using decisions based on summaries of category instances. The last two experiments assessed premise typicality as feature inference in classification learning of perceptual categories (Chapter 4). So, all of these experiments were assessments of category-based feature inference.

Chapter Two - Feature Inference Learning Compared to Classification Learning

2.1. General Introduction

There are at least two ways to learn about categories via feedback: classification learning and feature inference learning, as discussed in Chapter 1 (pp. 14-16). A key difference between classification and feature inference learning tasks is the presence of the category label as part of the stimulus in feature inference learning. This difference in available information suggests that classification and feature learning could result in fundamentally different category representations (see Anderson et al., 2002; Gelman & Markman, 1986; Johansen et al., 2015; Yamauchi & Markman, 1998; Yamauchi & Markman, 2000). Yamauchi and Markman (1998) proposed that feature inference learning promotes learning the internal structure of a category including prototypical features, consistent with a prototype representation and in contrast to classification learning. Johansen et al. (2015) found that the label has a larger impact on learning than other features because it is more salient. Adapting the hypotheses from Yamauchi and Markman (1998; 2000) in light of Johansen et al. (2015), I propose a label induced rule bias hypothesis: category labels in feature inference bias participants to use the labels to try to form rules. In contrast, classification learning does not result in such a bias due to the lack of the category labels as part of the stimuli. Note that this is not a hypothesis about what representation participants *definitely* use but rather a bias for *trying* to use a representation based on the category labels as explained in detail below.

The purpose of this research was to go back to an important starting point for category learning--Shepard et al. (1961)--and to re-evaluate these classic category structures, see Figure 1 (and described in Chapter 1, pp. 17-20) in terms of feature inference learning to assess the underlying category representation. The conceptual reason these category structures are important is that they represent an evaluation of the learnability of what are among the simplest, non-trivial categories. Given the conceptual importance of the Shepard et al. (1961) types and

prior research on potential differences between classification and feature inference learning, the aim of Experiments 1-2 was to compare the category representation for classification and feature inference learning of the classic types. The label bias hypothesis predicts that feature inference learning induces a tendency to form rules (and therefore a rule-based representation) based on the category labels starting with simple unidimensional rules and progressing to more complex rules if required. This can potentially be observed in terms of advantages for feature inference over classification learning wherever a label-based rule allows performance above chance e.g. unidimensional rules for Types I and V (Figure 1 in Chapter 1, p. 18). This is because the bias corresponds to participants trying to form a label-based rule first over other feature-based rules and in Types I and V these rules are diagnostic and semi-diagnostic respectively. In classification tasks there is no such bias as all stimulus features are roughly comparable in nature, potentially leading to a difference in the representation underlying classification and feature inference learning.

2.2. Experiment 1

Experiment 1 compared the learnability of a subset of the classic Shepard et al. (1961) types, Figure 1 (Chapter 1, p. 18); specifically Types I, II, V and VI by classification and feature inference. Not all the features in the category structure types can be unambiguously learned by feature inference. Consider Type I as shown in Table 1 for the rocket ship stimuli in Figure 2 and assume that the dreton category includes the four instances in the top four rows of Table 1. Suppose the participant is shown that an instance is a member of the category, 'dreton', has narrow wings and a small booster (A11_ in Table 1, Type I). If asked to infer what the length of the body band should be (A11?), there are two dreton category instances that have narrow wings and a small booster but one of them has a long body band (A111) and the other has a short body band (A110), so this feature inference cannot be accurately learned for this type. However, other feature inferences can be accurately learned, for example for Type I, knowing

that a stimulus is a dreton with a large booster and a long body band (A?11) only corresponds to one instance in the category structure, and thus its wing size can be unambiguously inferred as narrow (A111).

Table 1

Abstract category structures for each of the two learning conditions (classification and feature inference) and the four category structure types (I, II, V and VI) used in Experiment 1. Training phase trials at the top and testing phase trials at the bottom.

Classification Training Phase

Type I	*Type II*	*Type V*	*Type VI*
A 111	**A** 111	**A** 111	**A** 111
A 101	**A** 110	**A** 110	**A** 010
A 110	**A** 001	**A** 101	**A** 001
A 100	**A** 000	**A** 000	**A** 100
B 011	**B** 011	**B** 011	**B** 011
B 001	**B** 010	**B** 001	**B** 110
B 010	**B** 101	**B** 010	**B** 101
B 000	**B** 100	**B** 100	**B** 000

Feature Inference Training Phase

Type I	*Type II*	*Type V*	*Type VI*
A **111**	A **111**	A **111**	A **111**
A **101**	A **110**	A **110**	A **010**
A **110**	A **001**	A **101**	A **001**
A **100**	A **000**	A **000**	A **100**
B **011**	B **011**	B **011**	B **011**
B **001**	B **010**	B **001**	B **110**
B **010**	B **101**	B **010**	B **101**
B **000**	B **100**	B **100**	B **000**

Testing Phase

Type I	*Type II*	*Type V*	*Type VI*
A 111	**A** 111	**A** 111	**A** 111
A 101	**A** 110	**A** 110	**A** 010
A 110	**A** 001	**A** 101	**A** 001
A 100	**A** 000	**A** 000	**A** 100
B 011	**B** 011	**B** 011	**B** 011
B 001	**B** 010	**B** 001	**B** 110
B 010	**B** 101	**B** 010	**B** 101
B 000	**B** 100	**B** 100	**B** 000
A **111**	A **111**	A **111**	A **111**
A **101**	A **110**	A **110**	A **010**
A **110**	A **001**	A **101**	A **001**
A **100**	A **000**	A **000**	A **100**
B **011**	B **011**	B **011**	B **011**
B **001**	B **010**	B **001**	B **110**
B **010**	B **101**	B **010**	B **101**
B **000**	B **100**	B **100**	B **000**
A **1?1**	A **111**	A **110**	A **111**
A **1?0**	A **000**	A **000**	A **100**
B **0?1**	B **010**	B **010**	B **110**
B **0?0**	B **101**	B **100**	B **101**
A **11?**	A **11?**	A **101**	A **010**
A **10?**	A **00?**	A **000**	A **001**
B **01?**	B **01?**	B **001**	B **011**
B **00?**	B **10?**	B **100**	B **000**

Note. 'A' and 'B' refer to the two category labels and the subsequent three numbers refer to the three binary-valued feature dimensions and the feature values on those dimensions. Bolded features and question marks indicate what was queried for a given instance in a given condition. Bold features represent a correct answer and question marks indicate the lack of an unambiguous correct answer.

This experiment used Types I, II, V and VI because these allow all feature inferences on a single dimension, dimension one, to be unambiguously trained. In contrast, Types III and IV cannot be completely unambiguously trained by feature inferences on dimension one and consequently have not been evaluated here. So, all responses were on a single dimension for feature inference learning like they were in classification learning. It is also worth noting the classic finding that Types III, IV and V are equivalent in learning difficulty.

The stimuli in this experiment, the rocket ships in Figure 2, were used in preference to the classic colour/shape/size stimuli in Shepard et al. (1961) because those stimuli don't allow feature removability i.e. for individual features to be removed but the stimulus to still be presented. For example, you cannot remove the shape dimension from a 'large black triangle' as the colouring of the instance remains and needs to have a shape. In contrast, the rocket ship stimuli can be presented with individual features removed so that those features can be queried in terms of a feature inference learning task.

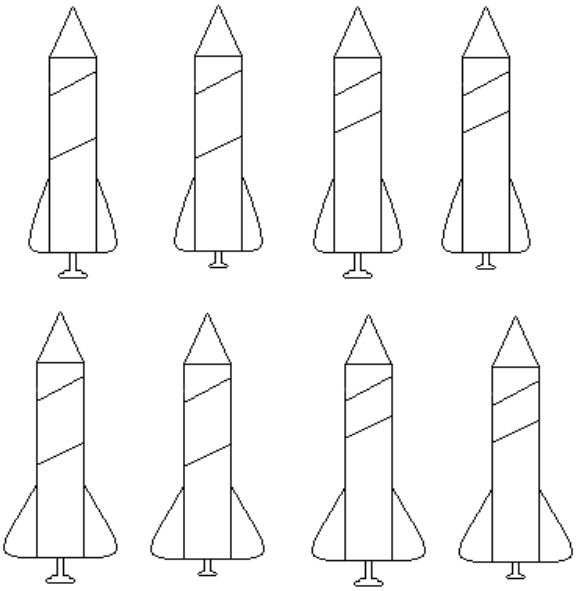

Figure 2. Set of eight rocket ship stimuli used in Experiment 1 composed of features from three dimensions--wing width, body band length and booster size.

2.2.1. Materials and Methods

2.2.1.1. Participants

120 Cardiff University students participated for either course credit or payment. 30 participants were trained on each of the category structure types (Table 1) with 15 in each learning condition: classification or feature inference.

2.2.1.2. Materials and Procedure

The abstract category structures corresponding to the four types--I, II, V and VI--and the specification of all training and testing trials, by type, are shown in Table 1. Each category type has eight instances equally split into two categories, see the top of Table 1. Each instance is composed of a category label, either A or B, and three binary-valued feature dimensions. Each column indicates one feature dimension with two feature values. Bold features and question marks indicate the feature that was queried on a trial. For example, the classification condition training item **A**111 indicates that the category label was queried and the feature inference training item A**1**11 indicates that the first perceptual feature dimension was queried. Bold features indicate the correct answer and question marks indicate that there was no unambiguous correct answer. Testing trials, at the bottom of Table 1, omitted feedback and included all classification and feature inference training items from both training conditions; a given participant was only trained on classification or feature inference but was tested on both in the testing phase. Testing also included a selection of feature inferences on the second and third dimensions as shown at the bottom of Table 1.

The eight rocket ship stimuli used, Figure 2, corresponded to the eight instances in each category structure type (Table 1). The rocket ships had features on three dimensions: wing width (wide or narrow), body band length (long or short) and booster size (large or small). The two categories of rocket ship were labelled ‘dreton’ and ‘rilbar’. The assignment of physical features, Figure 2, to abstract category features, Table 1, was randomized across participants.

The rocket ship stimuli were presented using DirectRT. Participants responded by choosing between two on-screen options via key press (the w-key response with a "left" sticker on the keyboard for the option on the left side of the screen and the p-key response with a "right" sticker for the option on the right). The response options were either the category labels or two different features. Participants completed 320 training trials consisting of the eight category instances in random order within a block for 40 blocks of training. Subsequent to training, participants were tested without feedback on a block of the classification training instances, followed by a block of the feature inference training instances and finally, a block of a selection of feature inferences on the second and third dimensions (see Table 1). The order of the testing trials was randomized within each block.

In the training phase, after each classification training response, feedback contained the full rocket ship stimulus, the words, 'correct' or 'incorrect' followed by 'This is a dreton' or 'This is a rilbar' depending on the correct answer. After each feature inference training trial there was feedback which contained the full rocket ship stimulus, the words, 'correct' or 'incorrect' followed by 'The correct answer is shown above'. Participants could look at each feedback screen for as long as they wanted and pressed the space bar to continue to the next trial.

Participants were given a cover story that they were visiting an alien solar system and needed to learn about the different types of rocket ships used by the aliens. They were told that they were going to be shown rocket ships and would need to make a choice between two responses. In the classification conditions, they were told that the category labels would be their response options, and in the feature inference conditions, they were told that two features would be their response options. They were made aware that they would have to guess initially but that they could learn the correct responses with practice. Following this were two practice

trials to ensure that participants were aware of which keyboard button related to which response. They then completed the learning phase followed by the testing phase.

2.2.1.3. Design

This was a between-subjects design with eight conditions: four structure types (Types I, II, V and VI as in Table 1) learned by either classification or feature inference. The key dependent variable was accuracy by block for each condition. I also report the proportion of participants whose performance was greater than a learning criterion as an alternative measure of learning.

2.2.2. Results

Asymptotic learning was fairly poor for all conditions except Type I as shown by the proportions of participants who reached a learning criterion of 75% correct in the last four blocks of training, Figure 3. Despite 40 blocks of training, only 13% of participants achieved the criterion in the Type VI feature inference learning condition. Learning in this experiment was substantially worse than a prior standard replication, Nosofsky et al. (1994), see Figure 4; that is, the average proportion correct in the last sixteen trials of classification training was worse for Type II ($t(14) = 3.9$, $p = 0.002$), Type V ($t(14) = 4.7$, $p < 0.001$) and Type VI ($t(18) = 6.2$, $p < 0.001$) in this experiment than in Nosofsky et al. (1994). Despite this, there is evidence that some learning occurred in all conditions as performance was significantly above chance for Type II ($t(14) = 4.0$, $p = 0.001$) and Type V ($t(14) = 3.0$, $p = 0.009$) and there was a marginal difference in the right direction for Type VI ($t(14) = 1.5$, $p = 0.144$. Note that the degrees of freedom for some of the prior and subsequent t-tests are adjusted degrees of freedom in the context of assuming unequal variances).

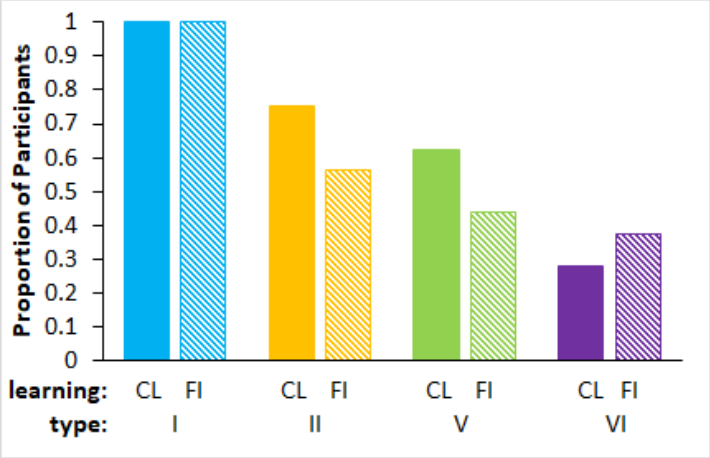

Figure 3. Proportion of the N = 15 participants in each learning condition and type from Experiment 1 who achieved the learning criterion (75% accuracy in the final four training blocks).

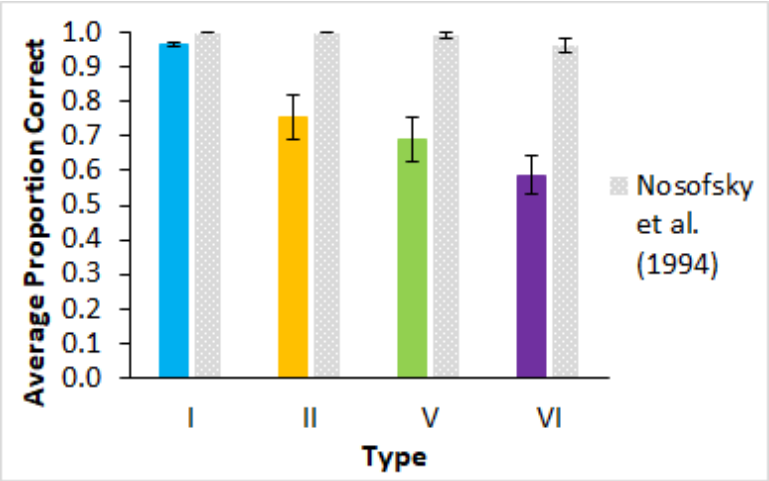

Figure 4. Accuracy in terms of proportion correct, averaged in the final learning phase block by type for the Experiment 1 classification conditions, coloured bars, and for data from Nosofsky et al. (1994), grey bars. Error bars show ±1 standard error.

For the classification learning task, Figure 5 left panel, average accuracy over all learning blocks, was higher for Type I than the next most accurate type, Type V ($t(19) = 5.5$, $p < 0.001$). There was no significant difference between Types V and II ($t(28) = 0.7$, $p = 0.501$), but average accuracy was significantly higher for Type II than Type VI ($t(15) = 2.7$, $p = 0.016$).

Overall, the results of the classification learning conditions replicate the classic difficulty ordering, though Types II and V were not clearly differentiated.

For the feature inference learning task, Figure 5 right panel, average accuracy was significantly higher for Type I than Type V ($t(21) = 12.0$, $p < 0.001$). Type V was not significantly different from Type II ($t(20) = 1.0$, $p = 0.339$) nor was Type II significantly different from Type VI ($t(21) = 1.5$, $p = 0.141$). Thus, feature inference learning only clearly replicated the classic finding in terms of Type I being the easiest with poor differentiation of Types II, V and VI.

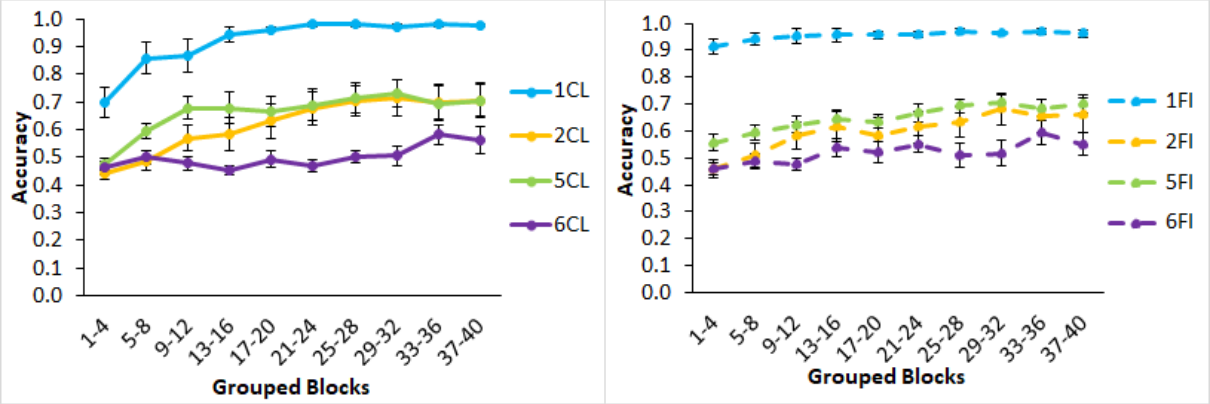

Figure 5. Average accuracy as proportion correct across groups of four training blocks by type (I, II, V and VI) and learning condition (CL = classification, FI = feature inference) in Experiment 1. The '1CL' header refers to the Type I classification condition, '1FI' refers to the Type I feature inference condition etc. Classification learning is displayed on the left, and feature inference learning is on the right. Error bars show ±1 standard error.

A power analysis is helpful to contextualize nonsignificant results for null differences between types, especially between Types II and VI feature inference learning, Figure 5. Power has been calculated in relation to the difference between the Types II and VI classification tasks in Nosofsky et al. (1994) as this is the smallest significant type difference in the present results. The effect size Cohen's *d* was specified using the difference of the average performance across blocks for the two conditions in Nosofsky et al. (1994) divided by a pooled estimate of the

standard deviation based on both, resulting in a *d* of 1.087. Assuming a significance level of alpha as 0.05, this effect size *d* and the number of participants in the Types II and VI feature inference conditions, the power (Howell, 2007) in the feature inference conditions to detect an effect this large was 0.85. (The classification learning data in this experiment is separate from the feature inference learning data, and the difference between Experiment 1 Types II and VI classification can also be used to calculate an effect size, $d = 0.996$, and non-post-hoc power for feature inference learning as 0.77). Thus, if the true difference between Types II and VI in feature inference was as large as the smallest significant type difference in the Experiment 1 classification learning results, this experiment had fairly high power to detect it. So, power analyses suggest that the null result in terms of a nonsignificant difference between Types II and VI in feature inference learning reasonably support a null conclusion of no difference between these two types by feature inference learning in contrast to the significant difference for classification. This indicates that performance in the Types II, V and VI feature inference conditions is very similar as there was likely enough power to detect a difference as big as those in classification.

Direct comparison of classification and feature inference learning by type shows that there were significantly higher accuracies for Type I feature inference than for Type I classification in the first four blocks of learning ($t(21) = 3.5$, $p = 0.002$). This superior performance in the feature inference condition supports the label bias hypothesis that feature inference learning induces a bias to evaluate the label-based unidimensional rule and classification does not due to the label not being present as part of the stimuli. The average accuracies for classification and feature inference learning across all learning blocks were not significantly different for Type II ($t(28) = 0.3$, $p = 0.737$), Type V ($t(21) = 0.3$, $p = 0.796$) or for Type VI ($t(28) = 0.8$, $p = 0.442$).

As a way of visualizing learning at the level of individual participants, the learning error diagrams in Figure 6 show responding on each learning trial for every participant arranged by learning type and condition. Each participant's responding is shown within a black rectangle outline, within this outline a trial is a single dot, black dots indicate response errors and white 'dots' represent correct answers on individual trials. Each column of dots is the response accuracy for each training item in a block in standardized order, as in Table 1, and each row of dots in a rectangle represents the accuracy for a given training item across all forty training blocks. Finally, participants in each condition have been arranged roughly by their learning performance, good learners toward the top and poor learners toward the bottom.

A key benefit of error diagrams is to be able to spot patterns in errors at the level of individual participants. Amongst these, perseverative suboptimal rule use can be seen, Figure 6, as systematic errors on particular instances i.e. as horizonal black lines. Such suboptimal rule use is most apparent for Type V feature inference where a label-based rule allowed 75% accuracy at the cost of consistent errors on the fourth and eighth instances in Table 1. Participants were operationally defined as using this suboptimal rule if their responding was consistent with this error pattern for over 15 blocks out of 40 (with a maximum allowed deviation from the pattern of one response per block). Approximately half of the participants in the Type V feature inference learning condition showed this pattern of responding, significantly more than in classification learning ($p = 0.018$, Fisher's exact test). Crucially, this occurred despite the existence of a corresponding suboptimal rule based on a single feature dimension also being available in classification learning. This suggests that participants were perseverating with a label-based rule in the feature inference learning condition which supports the label bias hypothesis.

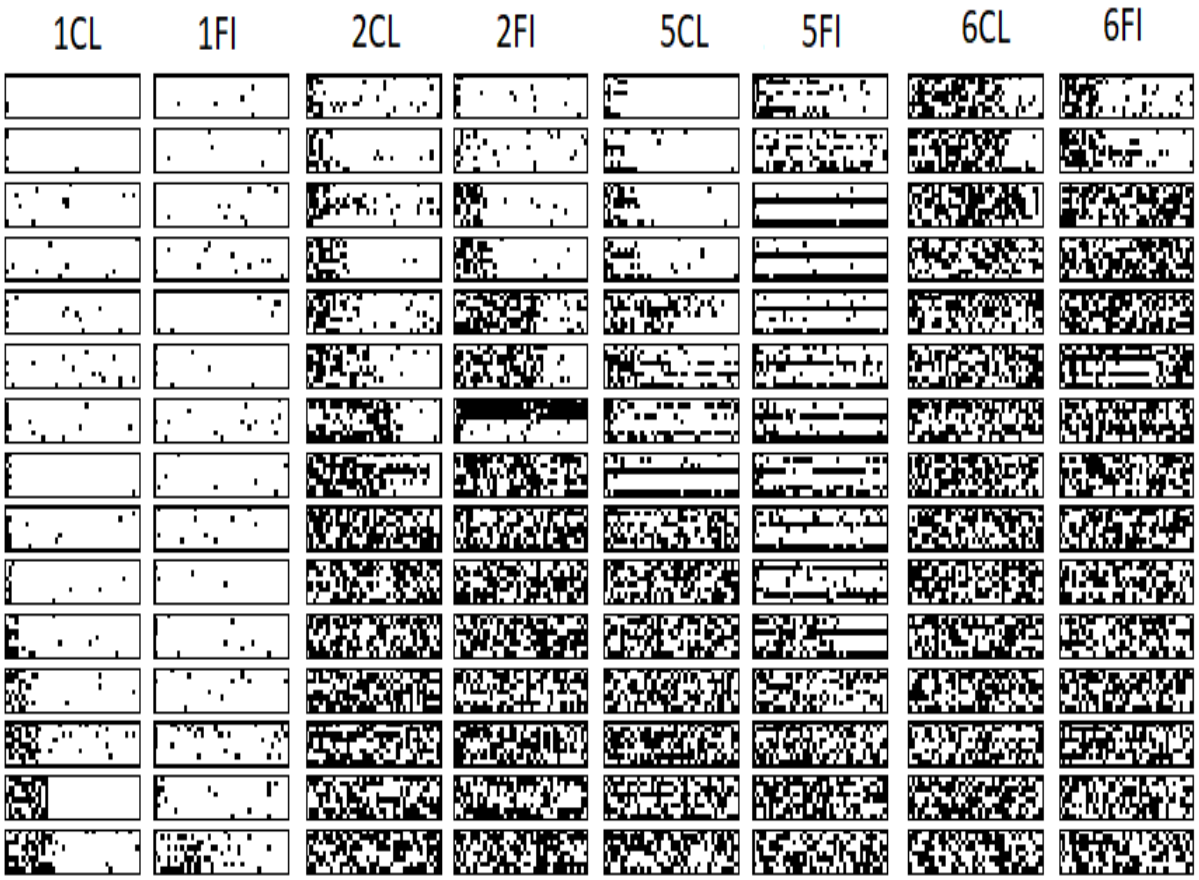

Figure 6. Error diagram panels showing the individual performance of each participant in Experiment 1 on every learning trial. Black dots = incorrect answers, white 'dots' = correct answers. Each row represents a single category instance, and the instances are ordered as in Table 1. Therefore, each row within a panel shows performance on one specific trial across the 40 learning blocks. Each column of panels represents a learning condition as indicated by the column headers. The '1CL' header refers to the Type I classification condition, '1FI' refers to the Type I feature inference condition etc.

Finally, the error diagrams show what appear to be rapid transitions from chance performance to high accuracy, seen as a change from the left (noise) to the right (white) of an individual panel. These potential rapid transitions are consistent with rule acquisition as finding a rule that gives optimal performance allows for rapid performance improvement.

2.2.3. Discussion

The results of this experiment support the label induced rule bias hypothesis which states that the category labels in feature inference learning bias participants to try to form label-based rules. This hypothesis is supported by the feature inference learning advantage for Type I, the presence of significantly more sub-optimal rule use for Type V feature inference compared to Type V classification and the similarity in the learning curves for Types II, V and VI feature inference.

In more detail, the label bias for Type I manifests as follows: in classification learning three unidimensional rules are all roughly equivalent, one for each feature dimension, with no clear basis for an initial preference between them, whereas feature inference has a single label-based unidimensional rule that is distinct from the other two feature based rules. Arguably, these differences arise out of a tendency to start with the label-based rule in feature inference, and therefore participants achieved perfect performance more rapidly in Type I. For Type I classification, learning occurred more slowly due to the lack of a clear basis for a preference between the three unidimensional rules. Some participants took longer than others to find the correct rule, and this greater variability resulted in classification participants, on average, taking longer to achieve perfect accuracy.

Further support for the label bias hypothesis comes from the perseveration of sub-optimal rule use in the Type V feature inference learning condition. A possible reason for this perseveration is in terms of difficulty as the relatively poor performance on the task overall compared to Nosofsky et al. (1994) and the interaction of this with the difference between classification and feature inference: the suboptimal label-based rule is easier to find in feature inference due to the bias and gives accurate enough performance to encourage participants to keep using the rule given the task difficulty.

Finally, the similarity of the learning curves for Types II, V and VI feature inference support the label bias as the bias is consistent with attempts to form label-based rules in feature inference conditions in contrast to classification. In Type I this leads to accurate performance early on in learning as is observed as the label-based rule is accurate. However, a bias for simple unidimensional rules for the harder types, doesn't allow optimal performance. In addition, even with the label bias, there are still multiple nonoptimal rules for the higher types involving the labels. So, a label bias for the higher types is less helpful for performance and may actually be harmful, manifesting in similar, relatively poor learning across the types.

The label bias hypothesis could be taken to imply that feature inference induces rule representation and that classification does not. However, the error diagrams suggest rapid transitions from chance performance to near perfect performance consistent with the use of rules in both classification and feature inference learning tasks for people who learned. The key difference is in terms of the label-based bias on rule formation in the feature inference learning conditions rather than a wholly different class of representation such as exemplars.

Perhaps the most surprising aspect of these results is the poorer learning of the classification conditions compared to Nosofsky et al. (1994): it is clear some learning occurred in all conditions, just not as much, see Figure 4. Methodologically, the classification learning conditions here were similar to standard replications with the key exception of the stimuli.

The current rocket ship stimuli are not unusual for the perceptual categorization paradigm where many prior studies have used rocket ships (see Craig & Lewandowsky, 2012; Johansen et al., 2015; Nosofsky et al., 1994; Palmeri, 1999; etc.). Also, different features were not visually hard to discriminate, see Figure 2. Kurtz et al. (2013) argued that the nameability of the feature values, the ease with which a feature can be given a verbal descriptor, influenced learnability; the nameability of Types II and IV impacts how well they are learned and can reverse the typical ordering for these types. The implication of nameability is in terms of the

implied interaction with the formation of verbal rules and their memorability. Minda, Desroches, and Church (2008) found that when a naming task was given to children prior to learning Type II, performance improved. To improve the learning in the classification conditions to be more consistent with prior replications, I adjusted the verbalizability of the stimuli in Experiment 2.

2.3. Experiment 2

To improve the nameability of the features on each dimension and reduce interference/confusability between dimensions, Experiment 2 changed the dimensions so that they were not all manipulations of size. In particular, the dimensions were colour, shape and size, as in the Shepard et al. (1961) stimuli, but applied to the rocket ship features, and the category labels were changed from two syllables to one syllable. In combination, these changes were intended to facilitate learning via more compact verbal rules. Importantly, rule use was assessed by asking participants to report what they saw and how they responded using qualitative questions at the end of the experiment. Finally, the feature inference feedback was amended slightly to include the category label, responding was via mouse rather than keyboard and the testing phase for both learning tasks was updated to include all possible feature inferences.

2.3.1. Materials and Methods

2.3.1.1. Participants

255 Cardiff University students participated for either course credit or payment. 64 participants completed each of category Types I, II, V and VI (Table 1) with 32 in each learning condition, classification or feature inference, except for Type II feature inference for which there were 31 participants due to an experimental error.

2.3.1.2. Materials and Procedure

The key change from Experiment 1 to Experiment 2 was the use of eight new rocket ship stimuli composed of three feature dimensions: body band colour (blue or green), cone shape (pointed or rounded) and wing size (wide or narrow) as shown in Figure 7. Additionally, the category labels were changed to 'thab' and 'lork' in an attempt to further aid the verbalizability of the stimuli and improve learning.

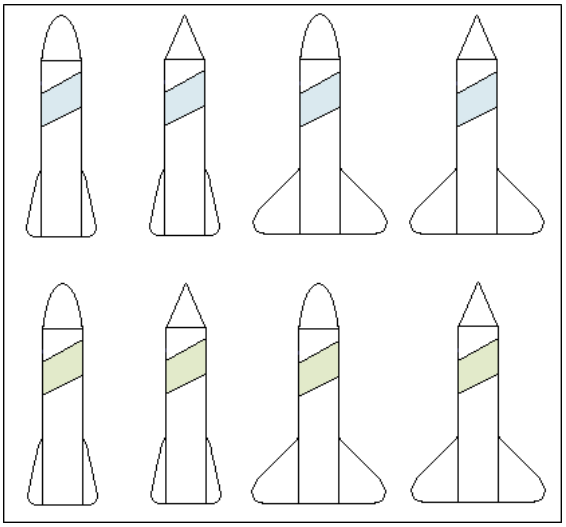

Figure 7. Rocket ship stimuli used in Experiment 2, composed of features on three stimulus dimensions: blue/green body band, pointed/rounded cone and wide/narrow wings.

Experiment 2 also had several minor adjustments. The feature inference feedback was adjusted to be identical to the classification feedback such that it included the label presented under the stimuli as well as the feedback, 'correct' or 'incorrect' and a note that the correct answer was shown at the top of the screen. The testing phase included all the original classification and feature inference items in Experiment 1, but feature inference trials were added to include all possible feature inferences on the training instances, see Appendix A. More importantly, at the end of the experiment, participants were given the following questions: 'Please write down the features you think changed between the different rocket ships.', 'Did you find a rule to help you learn the task? If so, please describe it briefly.' and 'If you did not find a rule what did you use/learn to help you do the task?' Participants responded via mouse

clicking images of the relevant feature/word rather than by pressing buttons on a keyboard and button pressing practice trials were eliminated. Participants moved on from each feedback screen by left-clicking the mouse rather than pressing the space bar. All other methodological details of this experiment were the same as in Experiment 1.

2.3.2. Results

The updated stimuli improved learning in all of the types, see Figure 8, with significant differences for Type I ($t(34) = 2.9$, $p = 0.006$), Type II ($t(91) = 3.6$, $p = 0.001$) and Type VI ($t(91) = 4.0$, $p < 0.001$), and at least a marginally significant improvement in Type V ($t(70) = 2.0$, $p = 0.046$). Overall, these results are consistent with more compact and less confusable verbal rules facilitating learning by being somewhat easier to use.

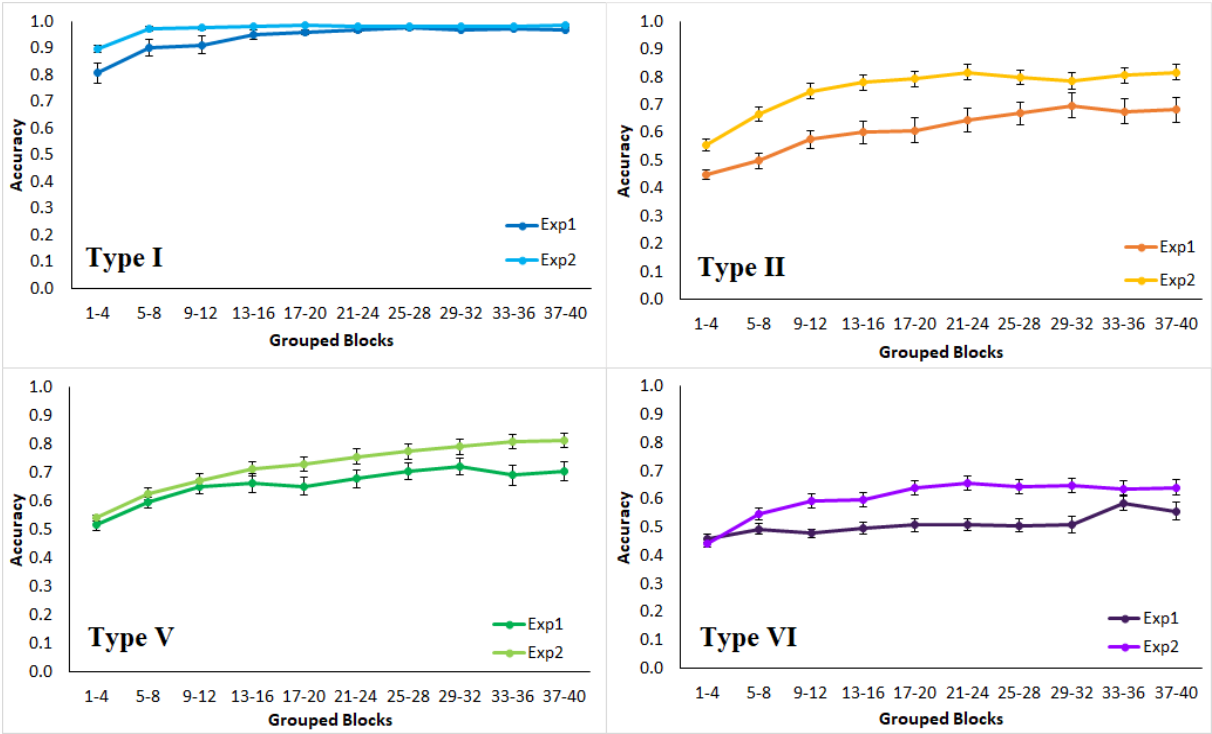

Figure 8. Comparison of average accuracy as proportion correct by type across groups of four training blocks between Experiments 1, dark lines, and 2, light lines. Error bars show ±1 standard error.

For average accuracy across all learning blocks in classification learning, Figure 9 left panel, accuracy was significantly higher for Type I than the next closest type, Type II ($t(33) = 6.4$, $p < 0.001$). There was no significant difference between Types II and V ($t(62) = 0.7$, $p = 0.492$) and accuracy for Type V was significantly higher than for Type VI ($t(62) = 4.4$, $p < 0.001$). Thus, the results of classification learning replicate the classic difficulty ordering, though Types II and V were again not clearly differentiated.

For average accuracy across all learning blocks in feature inference, Figure 9 right panel, Type I was significantly higher than Type II ($t(30) = 6.2$, $p < 0.001$) and there was no significant difference between Types II and V ($t(61) = 0.9$, $p = 0.390$) or between Types V and VI ($t(62) = 1.5$, $p = 0.128$).

Calculating power in relation to an effect size, $d = 0.607$, based on the difference between classification Types V and VI for the data presented by Nosofsky et al. (1994), the power to detect a difference this big between Types V and VI feature inference in the current experiment was 0.67. (The classification learning data in this experiment is separate from the feature inference learning data and the difference between Experiment 2 Types V and VI classification can also be used to calculate an effect size, $d = 1.096$, and non-post-hoc power for feature inference learning was 0.99. It is worth noting that the classification condition is a methodologically stronger comparison to the feature inference condition). Thus, feature inference learning replicated the ordering in Experiment 1 with Type I being the easiest and poor differentiation of Types II, V and VI, consistent with the label bias hypothesis.

Direct comparison of classification and feature inference learning by type shows that Type I feature inference had higher accuracy than classification across the first four learning blocks ($t(48) = 2.6$, $p = 0.013$). This replicates the findings of Experiment 1 and supports the label bias hypothesis. It is worth noting that a significant difference occurred despite the updated stimuli raising performance towards the ceiling. Additionally, when averaging over all

learning trials, there was higher accuracy for the Type VI feature inference participants than for the Type VI classification participants ($t(59) = 2.1$, $p = 0.039$). The average accuracies for classification and feature inference were not significantly different for both Type II ($t(61) = 0.1$, $p = 0.941$), and Type V ($t(62) = 0.4$, $p = 0.725$). These non-significant results, together with the previous power arguments, indicate the poorer differentiation of the learning curves in the feature inference learning conditions than in the classification learning conditions.

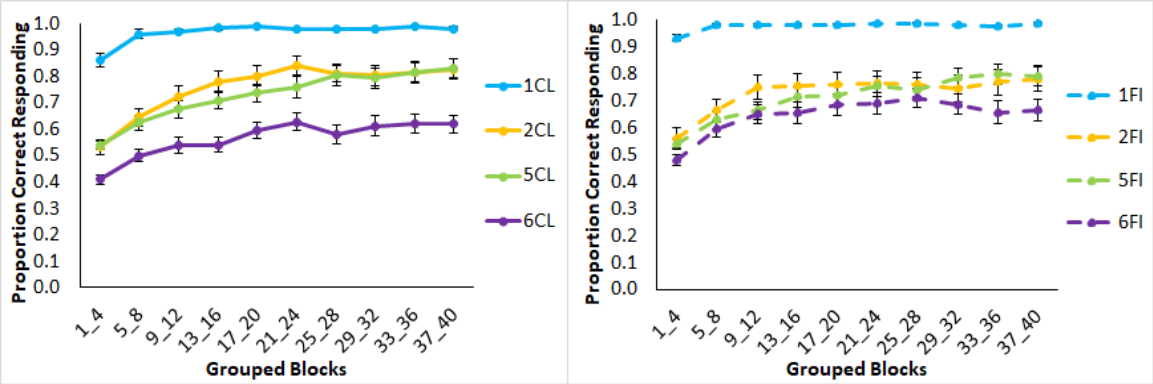

Figure 9. Average accuracy as proportion correct across groups of four training blocks by type (I, II, V and VI) and learning condition (CL = classification, FI = feature inference) in Experiment 2. The '1CL' header refers to the Type I classification condition, '1FI' refers to the Type I feature inference condition etc. Classification learning is displayed on the left, and feature inference learning on the right. Error bars show ±1 standard error.

As in Experiment 1, individual error diagrams, Figure 10, show rapid transitions from chance performance to high accuracy, consistent with the sudden acquisition of a rule. There was far less use of suboptimal rules, especially for Type V relative to the previous experiment, as might be expected due to the improved ease of learning given by the updated stimuli. This is consistent with more participants finding optimal rules, as supported by responses to the questions about learning strategy.

At the end of the experiment, participants described the stimuli they saw and how they used that information to learn the task. These qualitative data were used to assign participants

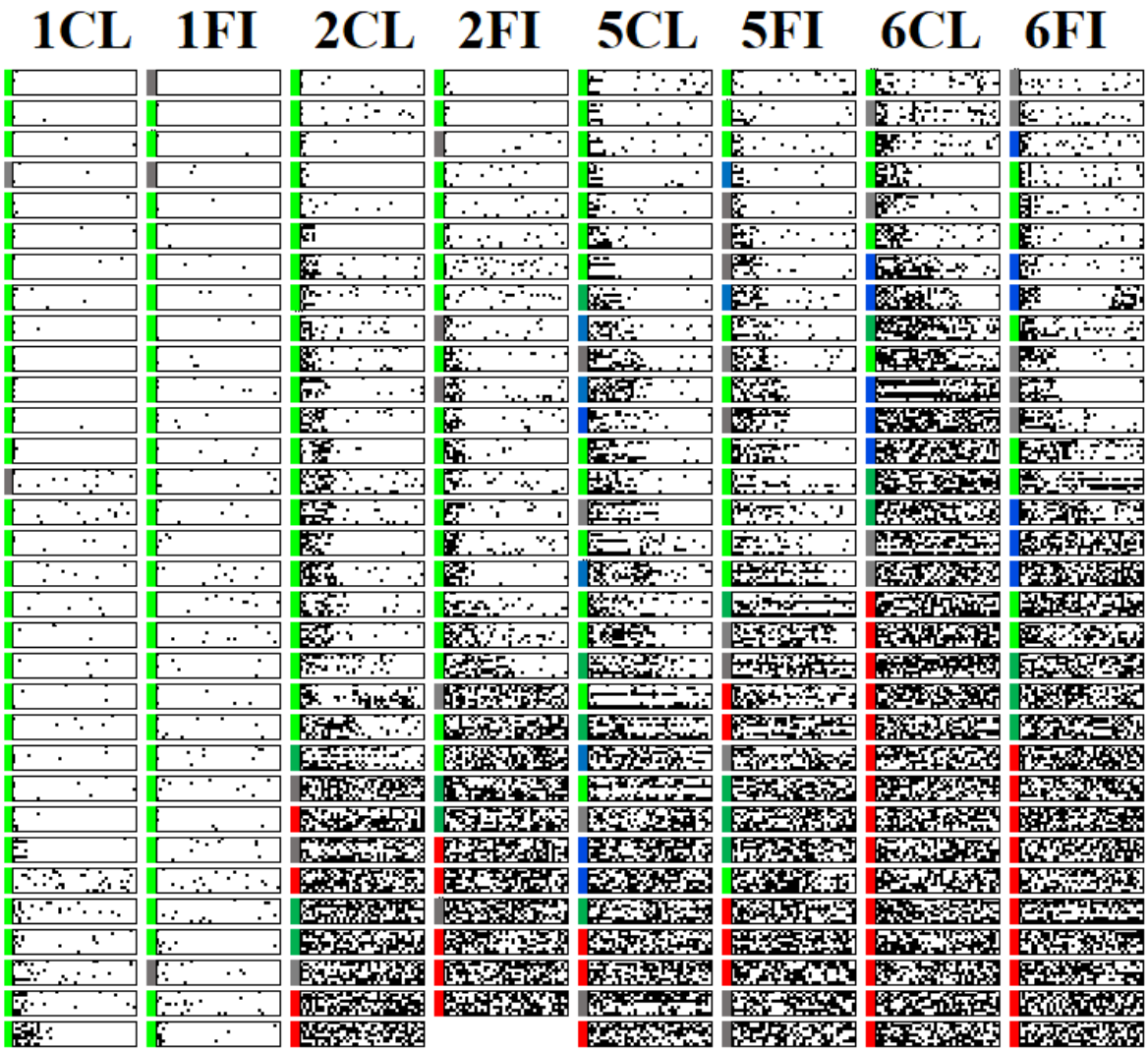

Figure 10. Panels showing the individual performance of each participant in Experiment 2 on every learning trial. Black dots = incorrect answers, white 'dots' = correct answers. Each row represents a single category instance, and the instances are ordered as in Table 1. Therefore, each row within a panel shows performance on one specific trial across the 40 learning blocks. Each column of panels represents a learning condition as indicated by the column headers. The '1CL' header refers to the Type I classification condition, '1FI' refers to the Type I feature inference condition etc. Blocks of colour to the left of each panel represent the learning strategy as inferred from the questions at the end of the experiment (light green = optimal rule, dark green = suboptimal rule, red = no rule/poor reported learning, blue = exemplars, grey = ambiguous).

to five groups in terms of those who used the optimal rule, suboptimal rules, no rule/poor reported learning, memorized exemplars or whose response was ambiguous. Participants were coded into the 'Optimal rule' group for Type I if they specified a first dimension, unidimensional rule, for Type II it was a configural rule based on the first two dimensions, for Type V it was a unidimensional rule with two, full instance exceptions and for Type VI it was the Odd-Even rule. Rules specified by participants were checked to confirm that they were optimal learning rules based on the specific stimuli a given participant actually saw. 'Suboptimal rules' were coded as any specification of a rule that was not the optimal rule for the condition a participant was in e.g. a unidimensional rule in Type II. Participants were coded into the 'No rule/poor reported learning' group if they specified that they had not been able to learn the task. The 'Exemplars' group was coded as a specification of more than two individual instances (so as to distinguish from optimal rule users in Type V specifying exceptions) or also specified as a mention of an instance memorization strategy. Finally, participants were coded into the 'Ambiguous' group if they gave responses that lacked sufficient information and/or were unclear.

The proportion of the participants attributed to each strategy for each learning condition, Figure 11 left panel, shows a predominance of optimal rule use, 67% (participants marked by light green patches, Figure 10) across Types I, II and V for participants in the error diagrams compared to the next highest strategy. As seen in the error diagrams for Type VI, the majority of participants did not learn the task. For all types, participants who reported that they did not learn, the red patches in Figure 10, were all participants in the error diagrams who clearly did not learn anything or who learned very little. When the participants who did not learn were removed, Figure 11 right panel, the proportion of optimal rule users averaged across all types was even higher, 79%. Despite the complexity of the Odd-Even rule in Type VI, roughly half of the participants who learned the task made a statement consistent with using this rule.

Although there were more participants who reported exemplar memorization (the blue patches in Figure 10) for the harder types, inclusion in this group was taken on participants' statement that they were using exemplars rather than a requirement to list all exemplars. In contrast, the attribution to the rule use group was based on the specification of a rule and a check that their rule would produce optimal performance.

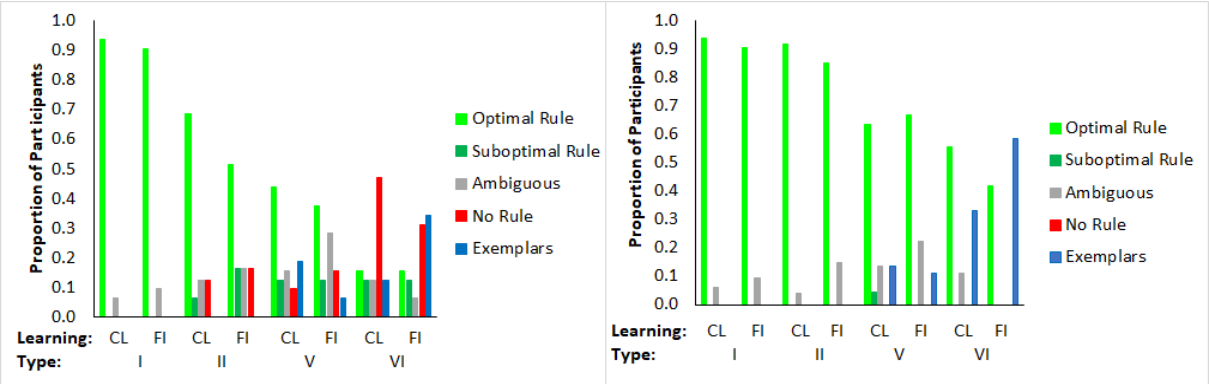

Figure 11. Proportion of participants who reported using each kind of representation for each type and learning condition. The left panel includes all participants and the right panel is for participants who met the learning criterion (where the number of participants who learned in each type and condition can been see in the error diagrams, Figure 10). Light green = optimal rule, dark green = suboptimal rule, grey = ambiguous, red = no rule/poor reported learning, blue = exemplars.

Participants in both learning conditions were tested on one block of classification items matching the classification training condition items and one block of feature inference items matching the feature inference training condition items, even though they were only trained on one of these learning tasks. For participants who met a 75% learning criterion in the last four learning blocks, the testing trial results, Figure 12, showed that decrements in performance between trained and untrained items were either tiny (classification) or not existent (feature inference). For example, one of the largest decrements was only 0.043 in Type II classification. This equates to one participant out of three making a single mistake across the eight untrained

testing items on average and shows very little decrement for the untrained trials. This is consistent with verbal rule use.

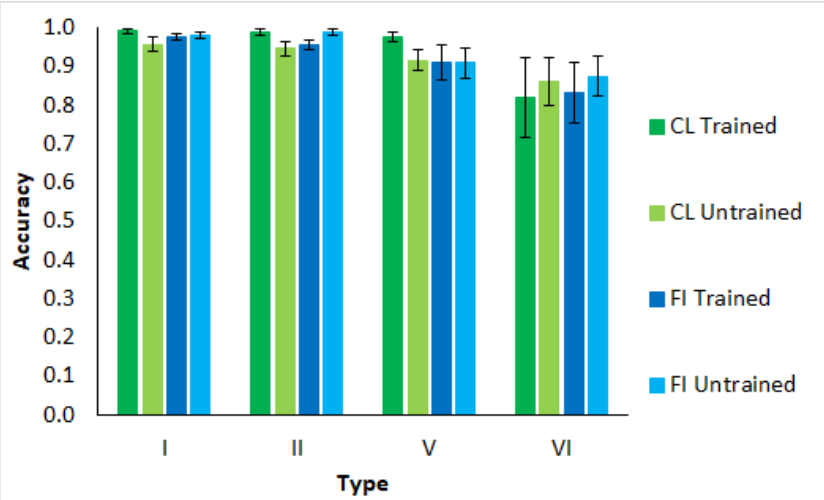

Figure 12. Average accuracy as proportion correct in the testing phase for each type, by prior learning condition and whether trials were previously trained or untrained for participants who met a 75% learning criterion in the last four learning blocks. Error bars show ±1 standard error.

2.3.3. Discussion

The label bias hypothesis was most directly supported by the Type I advantage. There was better performance in the Type I feature inference condition over the Type I classification condition early in training because the bias induced participants to try the correct rule more quickly. The hypothesis was also supported more subtly by the poor differentiation between Types II, V and VI in feature inference learning which is consistent with similar attempts to use label-based rules in all feature inference tasks even when these were not optimal rules.

As well as the support for the label bias hypothesis in terms of differences between the learning tasks, further support for this hypothesis comes from the pervasive use of rules in both learning tasks. This is supported by the rapid performance transitions in the error diagrams, the high accuracy on the untrained testing trials and the qualitative descriptions of rules that are accurate.

In more detail, the testing trial data showed essentially no decrements between performance on trained and untrained classification and feature inference trials. For example, a unidimensional rule in feature inference training might have been responding with wide wings for a thab and this would make it easy to do the untrained classification item that involved responding thab when shown a rocket ship with wide wings. The verbal rule explicitly contains the key elements of the stimuli and the category label, so it does not matter which is queried.

For the qualitative data, what people said they did was quite closely related to their actual performance. Bearing in mind that the strategy reported by 13% of all participants was ambiguous and a further 5% reported a strategy that was inconsistent with their performance, the strategy reported by 82% of participants was consistent with their learning performance either in terms of learning the task or saying that they didn't learn. For example, a participant in the Type VI classification condition who did not learn the task said, "No, didn't find a rule so I guessed each time." Another participant in the Type VI feature inference condition who did learn, stated, "If only one feature had changed, the top was the opposite to the previous rocket. If two changed, it was the same top as the previous rocket. If three changed, it was again the opposite." This is the Odd-Even rule. As it seems strange that participants could verbalize accurate rules if they were not using them, these data indicate the wide use of verbal rules across all types in both classification and feature inference learning.

Lastly, the learning improved significantly from Experiment 1 to Experiment 2, arguably due to the improvement of the nameability of the feature dimensions and values used for the stimuli and its subsequent impact on the ease with which the features could be used in verbal rules. However, it is important to note that learning was still poorer than in Nosofsky et al. (1994), see Figures 4 and 9, and without a direct comparison to the stimuli commonly used

in the Shepard et al. (1961) replications, the conclusions in terms of the nameability of the stimuli are limited due to other methodological differences.

Arguably, the stimuli used by Shepard et al. (1961) are specialized even by the standards of the perceptual learning paradigm in that their features can be described in a way that allows extremely compact verbal rules. In the classic stimuli, the noun descriptor of the object as a whole is treated as one of the features e.g. for a small, black triangle the name of the object overall is, 'triangle' however that is also one of the features. This contrasts with the rocket ship stimuli in which, 'rocket ship' is the name of the object but is not a feature value used to discriminate category instances. Additionally, the Shepard et al. (1961) stimuli are composed of features that refer to the instance as a whole and therefore do not require additional descriptors to discriminate between the feature dimensions. For example, with the size feature dimension, the Shepard et al. (1961) stimuli may have the value 'big' and that is sufficient to describe that feature value because it refers to the instance as a whole. With the rocket ship stimuli, the size dimension needs an extra descriptor, 'big booster' to indicate what is big. The learning based on the Shepard et al. (1961) stimuli therefore benefits from these advantages that allow rules to be specified very compactly. Experiment 3 directly compared the classic Shepard et al. (1961) stimuli to the rocket ship stimuli used in Experiment 2 in the context of a common methodology.

2.4. Experiment 3

The learning in Experiment 2 was not as good, Figure 9, as Nosofsky et al. (1994), Figure 4, the standard replication of Shepard et al. (1961). This is arguably due to the specialized nature of the stimuli used in Shepard et al. (1961) and its replications. Despite both sets of stimuli having the same feature dimensions of colour, shape and size, the Shepard et al. (1961) stimuli allow especially compact rules, e.g. **"black triangles, white circles group A**

else group B," in contrast to the rocket ship stimuli e.g. **"wide wings, pointed cone rockets and narrow wings, rounded cone rockets thab else lork."**

The purpose of this experiment was to contrast the classic stimuli with the rocket ship stimuli from Experiment 2 on Type II classification learning and compare the lengths of the implied learning rules from a qualitative question. I chose Type II as the configural rule involves four features per category and therefore the specification of this rule will include more descriptors than, for example, the unidimensional rule of Type I. This experiment did not include feature inference learning because the classic stimuli do not facilitate the feature removability needed for feature inference.

2.4.1. Materials and Methods

2.4.1.1. Participants

60 Cardiff University students participated for course credit or payment.

2.4.1.2. Materials and Procedure

The materials and procedure were identical to the Type II classification learning condition from Experiment 2 for the rocket ship stimuli condition, except for the removal of the feature inference testing items at the end. The second condition used the stimuli of Shepard et al. (1961) which included category labels (group A and group B) and dimensional variations of shape (triangle/circle), colour (black/white) and size (large/small). Note, the size dimension was scaled to be comparable to the overall size of the rocket ship stimuli.

2.4.2. Results

For the early learning blocks (1-4), Figure 13, learning was significantly faster for the classic stimuli than the rocket ship stimuli ($t(55) = 4.2$, $p < 0.001$). Individual error diagrams, Figure 14, also show this and replicate the rapid transitions from poor performance to high accuracy, indicative of rule use.

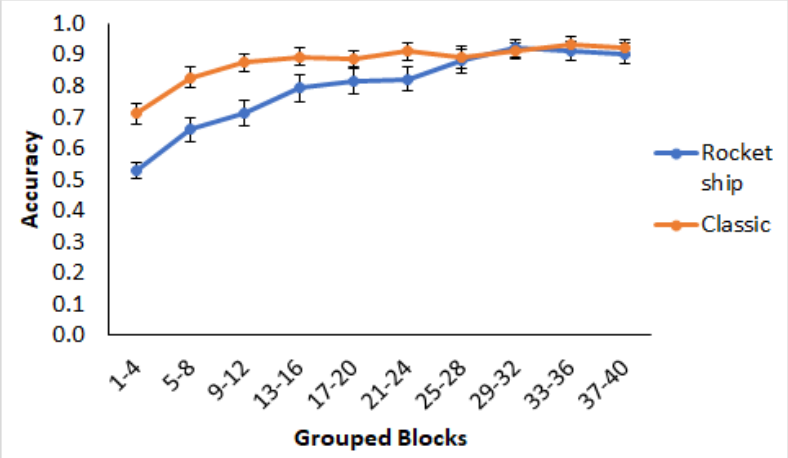

Figure 13. Average accuracy as proportion correct, averaged over four learning blocks for both the rocket ship stimuli condition and the classic stimuli condition. Error bars show ±1 standard error.

The key rule use query prompted participants to specify how they learned the task. Importantly, participants' full descriptions generally used more words than necessary to specify a rule and included comments not directly about their rule e.g. one participant stated, "Yes, leaving the mouse cursor on Group A I would select it if either an unfilled circle or filled triangle appeared. Otherwise I selected Group B then reset my cursor on Group A. I focused only on Group A by using true/false methods to switch and select Group B if necessary. Also, I repeated the words, "unfilled circle, filled triangle" in my head." From this I inferred the rule, "unfilled circle, filled triangle, Group A". Thus, data tabulation was in terms of a rule for one category with the other category implied to be instances that did not satisfy this rule, with additional comments and connectives removed, and only words which directly described features and category labels were included. The participants who did not learn were removed. There were significantly fewer words in the tabulated configural rules for the classic stimuli compared to the rocket ship stimuli, Figure 15, ($t(41) = 2.4$, $p = 0.021$).

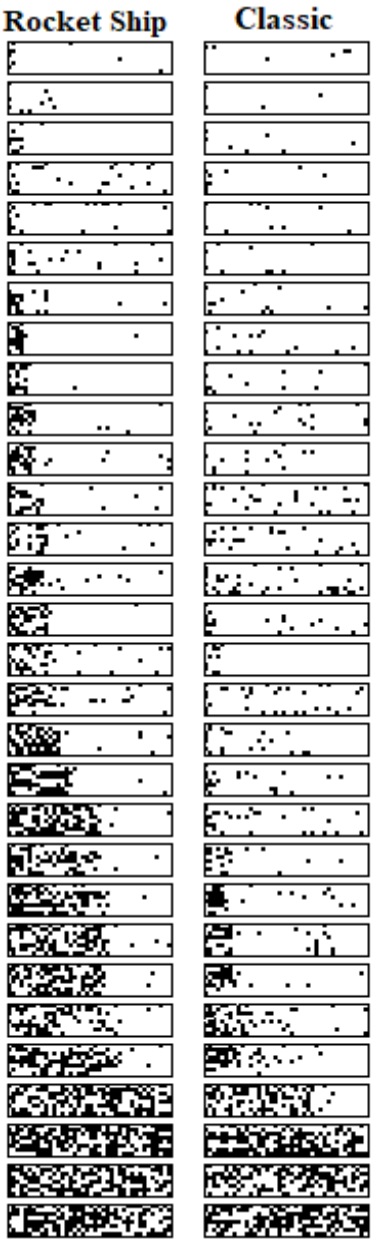

Figure 14. Error diagram panels showing the individual performance of each participant in Experiment 3 on every learning trial. Black dots = incorrect answers, white 'dots' = correct answers. Each row represents a single category instance, and the instances are ordered as in Table 1. Therefore, each row within a panel shows performance on one specific trial across the 40 learning blocks. Each column of panels represents a learning condition as indicated by the column headers. 'Rocket Ship' refers to the learning condition which used the Experiment 2 rocket ship stimuli and 'Classic' refers to the learning condition which used the Shepard et al. (1961) stimuli.

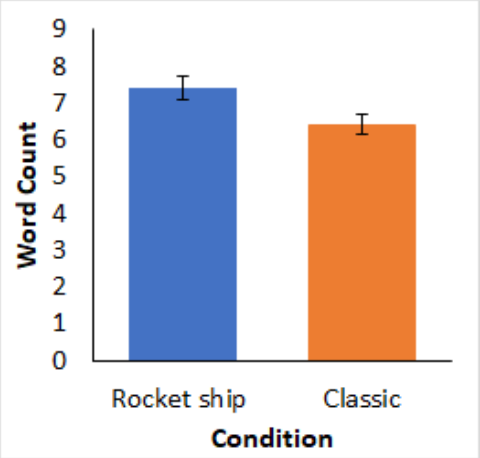

Figure 15. Average word count from the extraction of the specified verbal rule for the participants who achieved the learning criterion of greater than 75% correct over the last four learning blocks. Error bars show ±1 standard error.

The qualitative data was coded into an additional grouping in Experiment 3 that was not present in responding in Experiment 2. Participants were coded into the 'pattern of responding' group if they indicated that they responded with set answers regardless of what the stimuli were. Rule use, Figure 16, was high in both conditions with 70% of participants

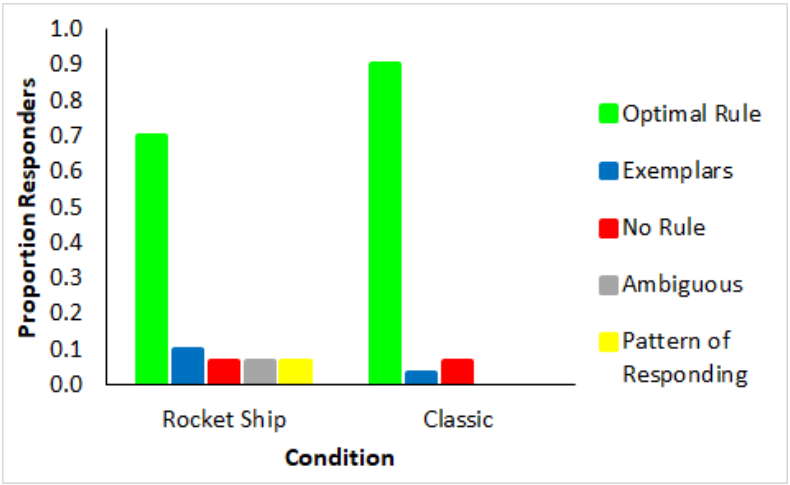

Figure 16. Proportion of all participants who reported using each kind of representation for both learning conditions in Experiment 3. Light green = optimal rule, blue = exemplars, red = no rule/poor reported learning, grey = ambiguous, yellow = pattern of responding.

reporting an accurate rule in the rocket ship stimuli condition and 90% of participants reporting an accurate rule when learning the classic stimuli. And the somewhat higher rule use for the classic stimuli was consistent with the better learning performance, Figure 13.

2.4.3. Discussion

Learning of the classic stimuli was better than the rocket ship stimuli, and the classic stimuli were described more compactly in verbal rules. This implies that the especial rule compactness of the classic stimuli influences learning performance of the Shepard et al. (1961) types; compact verbal rules facilitate learning.

2.5. General Discussion

Learning about categories by feature inference is plausible given the functional importance of feature inference in categorization. This research compared classification and feature inference learning of the classic category structures from Shepard et al. (1961; replicated and evaluated many times; Edmunds & Wills, 2016; Griffiths et al., 2008; Kruschke, 1992; Kurtz, 2007; Lewandowsky, 2011; Love, 2002; Love et al., 2004; Nosofsky et al., 1994; Rehder & Hoffman, 2005; Smith et al., 2004; Žauhar et al., 2016). These experiments provided support for the label-bias hypothesis i.e. a bias to try to use label-based rules in feature inference learning in contrast to classification learning. This manifested most directly in terms of Type I being learned faster by feature inference than by classification and by suboptimal rule use in Type V when the stimuli were hard to learn (Experiment 1). More subtly, this manifested in terms of less differentiation of the harder types for feature inference learning in contrast to the classic type ordering for classification learning.

Despite the support for the bias hypothesis, the results did not support a distinct kind of representation for classification versus feature inference learning; the results supported the preponderance of verbal rule representation for both in contrast to the conclusions of prior research (Anderson et al., 2002; Sweller & Hayes, 2010; Yamauchi & Markman, 1998; etc.).

Notwithstanding skepticism about self-report data, the qualitative data showed strong correspondence between what participants said about rules, what stimuli they saw and how well they learned. In particular a little under half of the participants who learned Type VI in Experiment 2 explicitly reported using the Odd-Even rule, a complex rule to articulate especially if this was not how they did the task. Further, the error diagrams, which showed individual participants' performance on individual trials (Figures 6, 10 and 14), demonstrated relatively rapid changes in performance from chance to high accuracy consistent with the acquisition of rules for many participants. The testing phase evaluated training instances from both classification and feature inference even though participants were only trained on one or the other. Testing trials showed little difference in accuracy on untrained responses (trials trained in the alternative learning condition) versus trained responses consistent with the use of rules, as a rule can be easily reversed in terms of stimulus and response.

The contrast in performance between Experiments 1 and 2 also supported the preponderance of rule-based representations; the implied greater difficulty of using the verbal rules on the stimuli in Experiment 1 corresponded to poorer performance but more suboptimal rule use especially for Type V. Changes to the stimuli to allow more compact verbal rules corresponded to better learning in Experiment 2, also shown by less suboptimal rule use in Type V. Experiment 3 directly compared the classic stimuli to the rocket ship stimuli from Experiment 2 and demonstrated the superior learnability of the classic stimuli, consistent with the argued relationship between rule compactness and learning difficulty.

While these results don't support a difference in the kind of representation for classification and feature inference learning (Anderson et al., 2002; Johansen & Kruschke, 2005; Sweller & Hayes, 2010; Yamauchi & Markman, 1998; etc.), the label-bias hypothesis is notably consistent with the spirit of the representational difference hypothesis from Yamauchi and Markman (1998), Anderson et al., (2002) etc. and with the importance of category labels

in category based decision-making (Gelman & Markman, 1986; Johansen et al., 2015; Yamauchi & Markman, 2000; etc.). Feature inference is plausibly less about the contrast between categories and more focused on the internal attributes of the category as centered on a conceptual label.

The fairly pervasive evidence for the use of rules for both classification and feature inference also has implications for various categorization models that have used Shepard et al. (1961) as a set of benchmarks. In particular, the fact that models such as ALCOVE (Kruschke, 1992) and SUSTAIN (Love et al., 2004) can account for the Shepard et al. (1961) results does not fit well with the prevalent evidence for rule representation in these tasks because the representations in these models are not explicitly rules, though it is worth noting that there was some evidence of exemplar use for some participants, especially for Type VI. The implications for dual-system models such as ATRIUM (Erickson & Kruschke, 1998) and COVIS (Ashby et al., 1998) are less clear because these models embody rule systems and similarity-based systems. Nevertheless, the evidence for rules here fits more comfortably with the fact that RULEX (Nosofsky et al., 1994) and DIVA (Kurtz, 2007) have been shown to be able to account for the classic Shepard et al. (1961) findings in terms of the representational assumptions embodied in these models.

The overarching aim of this thesis is to assess the nature of the representation underlying feature inference learning and decision-making. The present experiments assessed feature inference *learning* of the Shepard et al. (1961) structures and the results indicated that the dominant representation for these structures was rule based. However, these structures may be inordinately conducive to rules and as such would be expected to support this representation. And this may have especially been the case as the type seemingly most conducive to prototype representation, the Type IV family resemblance structure, wasn't included in the present experiment for reasons of methodological control (see p. 29). Given the ecological plausibility

of such structures, Experiments 4-8 assessed feature inference *decision-making* based on a variant of the family resemblance structure. The motivation for this was that prototypes are a plausible representational basis for categorical induction effects (Murphy, 2002), and a family resemblance structure should promote prototype representation. So, Experiments 4-8 in Chapters 3 and 4 were set up as a test of whether prototype representation underlies feature inference decision-making using perceptual categories that should maximally facilitate this assessment.

Chapter Three - Premise Typicality as Feature Inference Decision-Making in Perceptual Categories

3.1. General Introduction

3.1.1. Categorical Induction and Feature Inference

As described in Chapter 1, the categorical induction paradigm has a corpus of well-established empirical effects including premise typicality. However, the category representations underlying these effects are unclear, in part because the real-world categories commonly used in categorical induction tasks are so complex. In contrast, the perceptual category learning paradigm, with its highly controlled, simple stimuli, has fostered the development and evaluation of many well-specified formal models of category representation (Anderson, 1991; Kruschke, 1992; Love et al., 2004; Nosofsky et al., 1994; etc.). The similarities between the categorical induction and perceptual categorization paradigms suggest that the key categorical induction effects, particularly premise typicality, should occur via feature inference testing in the perceptual categorization paradigm, allowing an assessment of the representations underlying these effects and therefore feature inference using the many well specified representation models.

To investigate the premise typicality effect via feature inference in perceptual categorization, the following experiments used constructed categories that had two crucial properties necessary to be able to test premise typicality — a typicality structure and attached hidden features. First, the categories needed to contain instances with different levels of typicality. At least one instance needed to have a higher level of typicality than all others and another instance needed a lower level of typicality to correspond to a clear typical and atypical premise on which to base the test of premise typicality. Second there needed to be 'hidden' feature dimensions such that hidden features could be attached to the typical and atypical instances, corresponding to premise typicality statements in the categorical induction paradigm

where a known instance has a previously unknown property attached to it e.g. “Robins have property X”.

To assess premise typicality, the following experiments used decision-making tasks based on a category summary of simple rocket ship stimuli in two constructed categories, Figure 17. The rocket ships varied in terms of the width of the wings (wide or narrow), the body band length (long or short), the booster size (large or small) and the cone shape (pointed or rounded) and were in two categories, ‘dreton’ or ‘rilbar’, that were designed to have a typicality structure based on family resemblance.

The family resemblance structure and its variants have been commonly used in perceptual categorization learning because real-world categories tend to have family resemblance structures (see Love, 2002; Markman & Maddox, 2003; Minda et al., 2008; Rosch & Mervis, 1975; Ward, Vela, & Hass, 1990; etc.). The family resemblance structure in the following experiments was a constructed category structure with a strong typicality gradient which included a prototype, consisting of all typical category features, a set of instances that varied from the prototype by one atypical feature and a very atypical instance that varied from the prototype by having two atypical features, Table 2. So, for example, the dreton category prototype, the rocket outlined in green in Figure 17, had features typical of a dreton, in this case a long body band, small booster, pointed cone and narrow wings. The atypical instance, the rocket outlined in red in Figure 17, had two features typical of the dreton category, a long body band and small booster, and two atypical features, a rounded cone and wide wings. So, this category structure has the typicality gradient necessary for testing premise typicality.

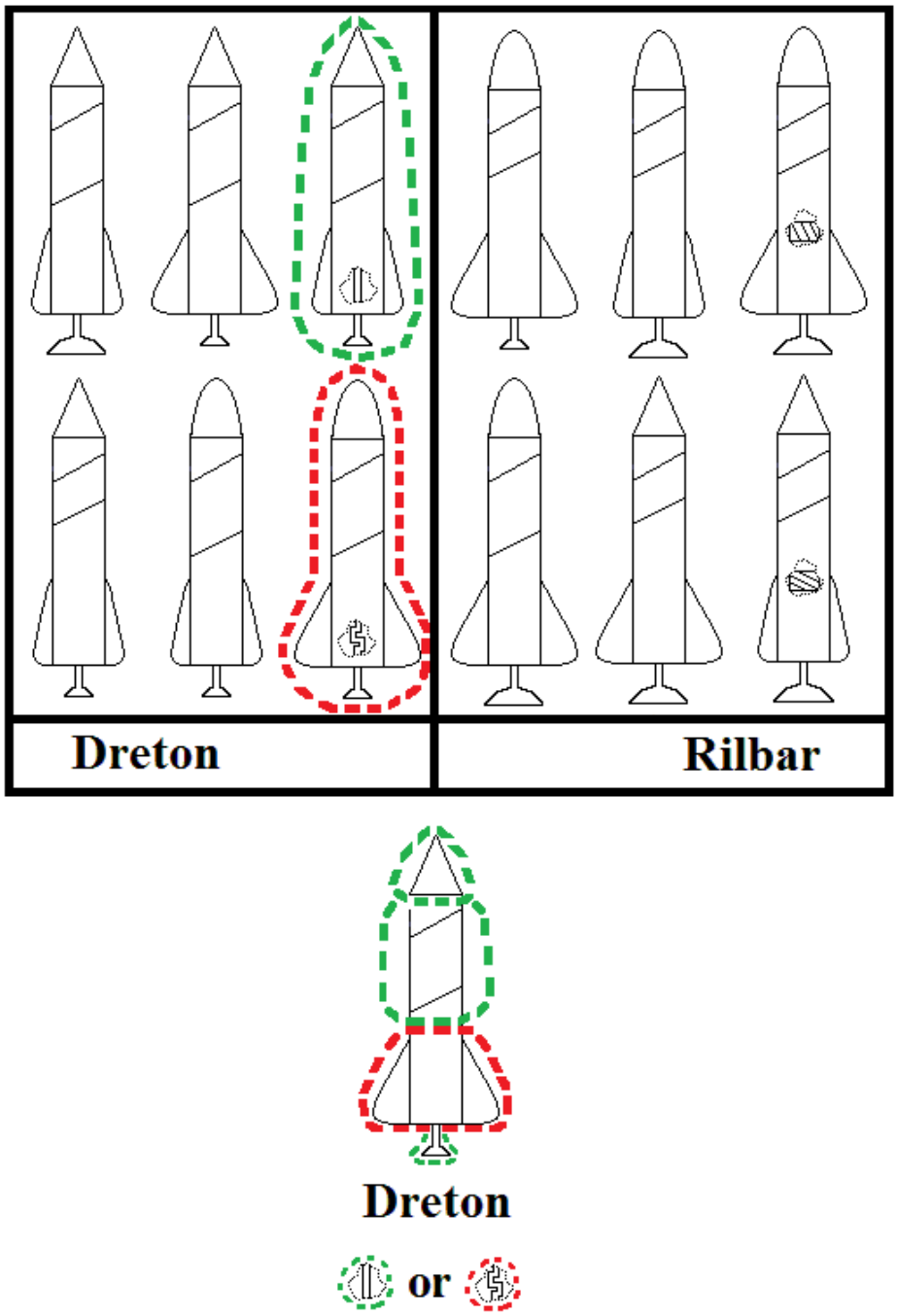

Figure 17. An illustration of a sample trial used in Experiment 4. The individual rocket ship at the bottom of the figure is a premise typicality testing trial including a rocket without a hidden feature, presented with its category label and two hidden feature response options. Typical features/instances are indicated by green dashed outlines and atypical features/instances by red dashed outlines, added for explanatory purposes only. Participants did not see these coloured outlines.

In the abstract family resemblance category structure for these experiments, Table 2, the typicality of a feature for a given category varies depending on how common that feature is within a category. In the table, each row specifies a specific category instance with six instances each for category A and B. For category A the most common value on every dimension is feature 1 and for category B the most common value is 3. Therefore, these values are the typical features whereas 3 for category A and 1 for category B are the atypical features. For example, in reference to Figure 17, a typical 1 feature for category A, 'dreton' is the small booster and the typical 3 feature for category B, 'rilbar' is the large booster. Conversely, the atypical 3 feature for the dreton category is the large booster and the atypical 1 feature for the rilbar category is the small booster. The other non-hidden feature dimensions were structured similarly. The 1 and 3 values on each dimension represent the two possible values each feature dimension could take: wide/narrow wings, long/short body band, large/small booster and pointed/rounded cone shape.

In Table 2, the hidden features are shown on hidden feature dimensions one and two with the letters V, X, Y and Z referring to physical features: straight and curved pipes and vertically and horizontally lined boxes, see Figure 17. Each binary-valued hidden feature dimension, pipes or boxes, occurred in only one category, with one hidden feature attached to the typical instance and the other hidden feature attached to the atypical instance. For example, the dreton category might have had the two pipe features associated with the typical and atypical instances, while the rilbar category might have had the two box features associated with its typical and atypical instances. A feature inference task tested premise typicality with the structure in Table 2 by attaching hidden features to the prototype (typical) and the atypical instances for each category and then querying which of these hidden features should be attached to a novel instance that did not (yet) show a hidden feature attached. The dotted cut-

Table 2.

The abstract category structure and key test cases for Experiments 4-8.

	Category	*Dimension 1*	*Dimension 2*	*Dimension 3*	*Dimension 4*	*Hidden Dimension 1*	*Hidden Dimension 2*
Classification	A	3	1	1	1	–	–
	A	1	3	1	1	–	–
	A	1	1	3	1	–	–
	A	1	1	1	3	–	–
	A	1	1	1	1	V	–
	A	3	1	1	3	X	–
	B	1	3	3	3	–	–
	B	3	1	3	3	–	–
	B	3	3	1	3	–	–
	B	3	3	3	1	–	–
	B	3	3	3	3	–	Y
	B	1	3	3	1	–	Z
Generalized Premise Typicality	A	1	1	3	3	–	–
	B	3	3	1	1	–	–
Ordinary Premise Typicality	A	3	1	1	1	–	–
	B	1	3	3	3	–	–
Premise Conclusion Similarity	A	1	3	1	1	–	–
	B	3	1	3	3	–	–
Blank Feature Inferences	A	–	–	–	–	–	–
	B	–	–	–	–	–	–

Note. The full abstract specification of all testing trials is in Appendix B. Colour coding refers to the typicality structure (green = typical, red = atypical, yellow = ordinary category instances). Dashes indicate the absence of a feature on a given dimension, see main text for an explanation of the testing trials.

out surrounding the hidden feature was intended to convey their hidden status by allowing participants to 'see into' the typical and atypical rocket ships whilst suggesting that the other rocket ships might have these features but that they are currently hidden due to the lack of a cut-out. An example of a premise typicality testing trial can be seen at the bottom of Figure 17.

This test instance has three features in common with the typical instance, a pointed cone, long body band and small booster. The test instance also has three features in common with the atypical instance, a long body band, wide wings and a small booster. Thus, the test instance was equally similar to the typical and atypical instances. The 'straight pipe' hidden feature was attached to the typical instance of the dreton category and the 'curved pipe' hidden feature was attached to the atypical instance. The key test of premise typicality was a new category instance that was equally similar to the typical and atypical instances, presented with its category label and the straight and curved pipes as the response options. So, a premise typicality effect in this paradigm corresponds to a preference for the feature associated with the typical instance, e.g. the straight pipe, over the atypical instance, the curved piped.

To summarize, Figure 18 shows the mapping between a standard premise typicality test in the categorical induction paradigm and the corresponding test in the perceptual categorization task used in the following experiments. The typical premise instance, 'a robin', shown on the top left of Figure 18, maps onto the typical rocket ship instance that is shown at the bottom of the left-hand column. The atypical premise, 'a penguin', shown on the top of the middle column in Figure 18, maps onto the atypical rocket ship instance. The conclusion instance, 'a crow', shown on the top of the right column, maps onto the test rocket ship instance which is equally similar to the typical and atypical rocket ship instances. The 'sesamoid bone' feature maps onto the hidden feature a rocket is shown to have e.g. a straight or curved pipe. There are two hidden feature response options for the rocket ship stimuli to allow a contrast between a hidden feature associated with the typical instance, 'the straight pipe' and a hidden feature associated with the atypical instance, 'the curved pipe'. This contrast between two hidden features is not necessary in the classic categorical induction paradigm as responses are judgements of argument strength rather than binary choices. However, in perceptual categorization, feature inference responding requires different hidden features be attached to

the typical and atypical instances so as to assess a preference for generalizing the hidden feature associated with the typical instance over the hidden feature associated with the atypical instance.

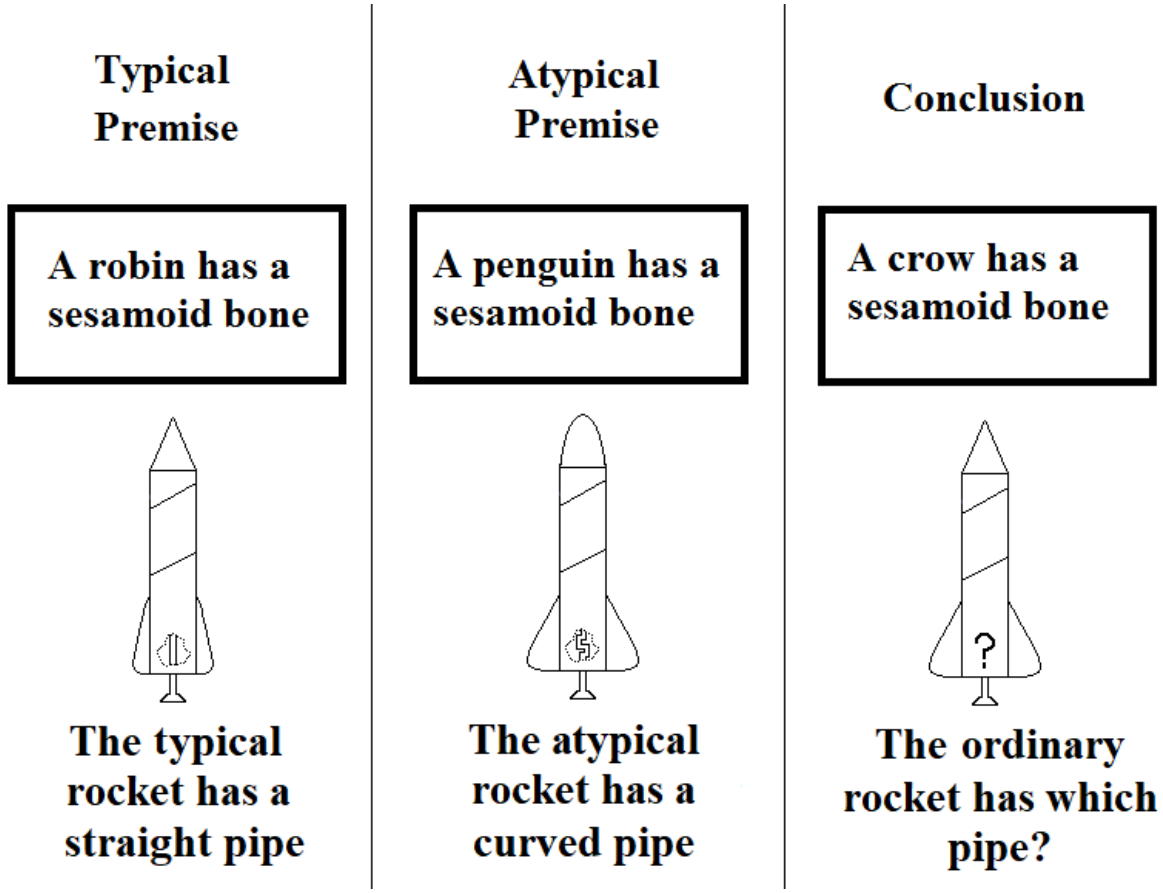

Figure 18. A summary of the mapping between premise typicality in the classic categorical induction paradigm as linguistic descriptions and in the perceptual categorization paradigm as perceptual rocket ships. Note that the participants did not see the phrases, 'The typical rocket has a straight pipe', 'The atypical rocket has a curved pipe' or 'The ordinary rocket has which pipe?' on the screen in the course of the experiment; these were added to the figure for explanatory purposes only.

3.1.2. Key Testing Trials

3.1.2.1. Tests of Premise Typicality

These experiments measured the key effect of premise typicality in two ways, see Table 2. The first, 'generalized' premise typicality, was based on instances not presented in the category summary such as A1133?, where '?' indicates a label or feature being queried at test and for this instance indicates that the hidden feature was being queried (see Table 3 for all testing trials of this type). This instance has two features in common with the typical instance and two features in common with the atypical instance, making it equally similar to both. Second, a test of 'ordinary' premise typicality was based on instances in the category summary, such as A3111? in Table 2, that had three features in common with both the typical and atypical instances, (see Table 3 for all testing trials of this type). So, as previously stated, a premise typicality effect corresponds to a preference for the hidden feature attached to the typical instance over the hidden feature attached to the atypical instance in both premise typicality trial types.

3.1.2.2. Unambiguous Testing Trials with Clear Correct Answers

Three sets of testing trials assessed participants' engagement with the task via questions with clear correct answers in the category summary. The classification testing trials included all 12 instances present in the category summary, see Figure 17 and Table 2. On each classification testing trial, one instance occurred individually underneath the summary with the response options as the categories dreton and rilbar. Incorrect answers to these trials imply participants likely weren't fully attending to the task. Further, the classification trials included typical, atypical and ordinary category instances which allowed for tests of typicality. A typicality effect should correspond to an appreciation of the typicality structure of the category as, for example, shown by greater accuracy on the typical instances than the atypical instances.

The hidden feature inference trials presented the typical and atypical instances without hidden features, and participants inferred those hidden features based on the 'correct answer' in the category summary, see Table 2. For example, in Figure 17, the green dotted instance in the dreton category would have been presented underneath the category summary without a cut-out showing the straight pipe, but the pipe would still be visible on the matching rocket in the summary. The straight and curved pipe would be the response options. This tested the attachment of the hidden features to the typical and atypical instances, a crucial property for the assessment of premise typicality.

Table 3.

The abstract structure for all testing trials in Experiment 4.

Abstract Structure	Trial Type	Testing Trials
A 3111_ _ A 1311_ _ A 1131_ _ A 1113_ _ A 1111V_ A 3113 X_ B 1333_ _ B 3133_ _ B 3313_ _ B 3331_ _ B 3333_Y B 1331_Z	Block 1 Classification	A3111_ _ A1311_ _ A1131_ _ A1113_ _ A1111_ _ A3113_ _ B1333_ _ B3133_ _ B3313_ _ B3331_ _ B3333_ _ B1331_ _
	Block 2 Generalized Classification	?1133_ _ ?3311_ _ ?1313_ _ ?3131_ _
	Block 3 Generalized Premise Typicality	A1133?_ B3311_? A1313?_ B3131_?
	Block 4 Original Premise Typicality	A3111?_ A1113?_ B1333_? B3331_?
	Block 5 Premise Conclusion Similarity	A1311?_ A1131?_ B3133_? B3313_?
	Block 6 Hidden Feature Inference	A1111V_ A3113X_ B3333_Y B1331_Z

Continued

Block 7 Ambiguous	A?111_ _ B?333_ _ A111?_ _ B333?_ _
Block 8 Exception Feature Inference	A3113X_ A3113X_ B1331_Y B1331_Y
Block 9 Label vs Feature	A3?33_Y B1?11_V
Block 10 Continuous Generalization	A2112?_ B1221_? A0110?_ B4114_? A1221?_ B3223_? A1001?_ B3443_?
Block 11 Blank Feature Inference	A_ _ _ _?_ B_ _ _ _ _?
Block 12 Label vs Hidden Features	A_?_ _ _Y A_ _?_ _Z B_?_ _V_ B_ _?_X_
Block 13 Premise Diversity	A2212?_ B2232_?
Block 14 The Inclusion Fallacy	A_ _ _ _?_ A3003?_ B_ _ _ _ _? B1441_?

Note. Question marks indicate no answer in the category structure, red labels/features were queried. See main text and Appendix C for all testing block descriptions.

Finally, the exception feature inference testing trials had only one exact match in the category summary and tested for correct exception feature inferences on the first or fourth feature dimensions, see Table 3, i.e. they tested the inference of atypical/non-prototypical features. For example, in Figure 17, the red dotted instance in the dreton category would have been presented below the category summary with the wide wings missing and the curved pipe hidden feature present. The response options would have been wide wings and narrow wings. This tested the ability to correctly attach the non-hidden features to the atypical category instances.

3.1.2.3. Premise Typicality *Like* Effects

Osherson et al. (1990) specified premise conclusion similarity effects as distinct from premise typicality effects. Similarity effects occur in terms of higher argument strength ratings when the premise and conclusion are very similar than when the premise and conclusion are dissimilar. The present experiments tested premise conclusion similarity using trials where the conclusion (the tested instance) was more similar to the typical instance than to the atypical instance, and participants chose between the hidden feature attached to the typical versus atypical instance. For example, the testing trial A1311 in Table 2, has three features in common with the typical instance for category A and only one feature in common with the atypical instance (see Table 3 for all testing trials of this type). A preference for the typical hidden feature on this test shows a premise typicality *like* effect that nonetheless is confounded with similarity because it can be based on similarity rather than typicality.

Finally, Blank Feature Inference tests included instances with no features and only a category label, see Table 2. These could also show a premise typicality *like* effect in terms of preferential responding with the typical hidden feature over the atypical in the absence of specific feature information.

3.1.3. Experiment Overview

The following three experiments all used the same category structure and testing trial types, Table 3, to assess categorical induction in category summary-based decision-making, most importantly, premise typicality. Experiment 4 investigated premise typicality but only included a limited number of trials with clear correct answers before the crucial test of premise typicality. Experiment 5 mixed in trials with clear correct answers to encourage participants to carefully attend to the instances in the category summary, especially the instances with hidden features. To further enhance attention to the summary, Experiment 6 added initial training trials with corrective feedback and asked participants to categorize the ordinary instances in the category and infer each non-hidden feature from those instances using the category summary. These experiments followed a progression of decision-making tasks that promoted increased use of and engagement with the category summary in terms of appreciating the typicality structure and attaching the hidden features to instances in that structure, both necessary prerequisites for assessing premise typicality.

3.2. Experiment 4

3.2.1. Introduction

Category-based decision-making has been widely evaluated using summary presentations of constructed categories (Griffiths et al., 2012; Johansen et al., 2015; Murphy & Ross, 1994; Murphy & Ross, 2010; Yamauchi & Yu, 2008; etc.). For example, Murphy and Ross (1994) used a decision-making task in which they presented sets of children's drawings and participants judged which child drew a test item and how likely the drawing was to have a certain feature. They found that participants tended to use only the category that the instance was most likely to have come from when making predictions; they don't tend to use multiple categories to inform their decisions. As a further example, Griffiths et al. (2012) used a category summary of alien bug stimuli to assess feature inference and they found that participants

preferred to use information about the features of the category instances when reasoning. The present experiment used a summary presentation methodology similar to that in Murphy and Ross (1994) and Griffiths et al. (2012) to evaluate categorical induction effects, particularly premise typicality. This was to allow for an assessment of the underlying category representation of these effects as elicited by feature inference testing.

3.2.2. Materials and Methods

3.2.2.1. Participants

40 Cardiff University students participated for course credit or payment; however, the reported data includes 37 participants as 3 data sets were lost due to experimental error.

3.2.2.2. Materials and Procedure

In total there were 16 rocket ship stimuli that varied on four binary valued dimensions, see Figure 19. These dimensions were wing size (wide/narrow), body band size (long/short), booster size (large/small) and nose cone shape (pointed/rounded). In addition, there were two hidden feature dimensions, pipes and boxes, each with two different features as, for example, shown in Figure 17.

The assignment of the four physical stimulus dimensions in Figure 19 to the four abstract dimensions composing the 16 abstract category instances, see Table 3, was randomly allocated for each participant from a set of possible assignments, as was the assignment of the physical features comprising the two hidden features dimensions and their values. Similarly, the category labels dreton and rilbar were also randomly allocated to the two abstract categories, A and B, in Table 2. The ordering of the trials was randomized within blocks.

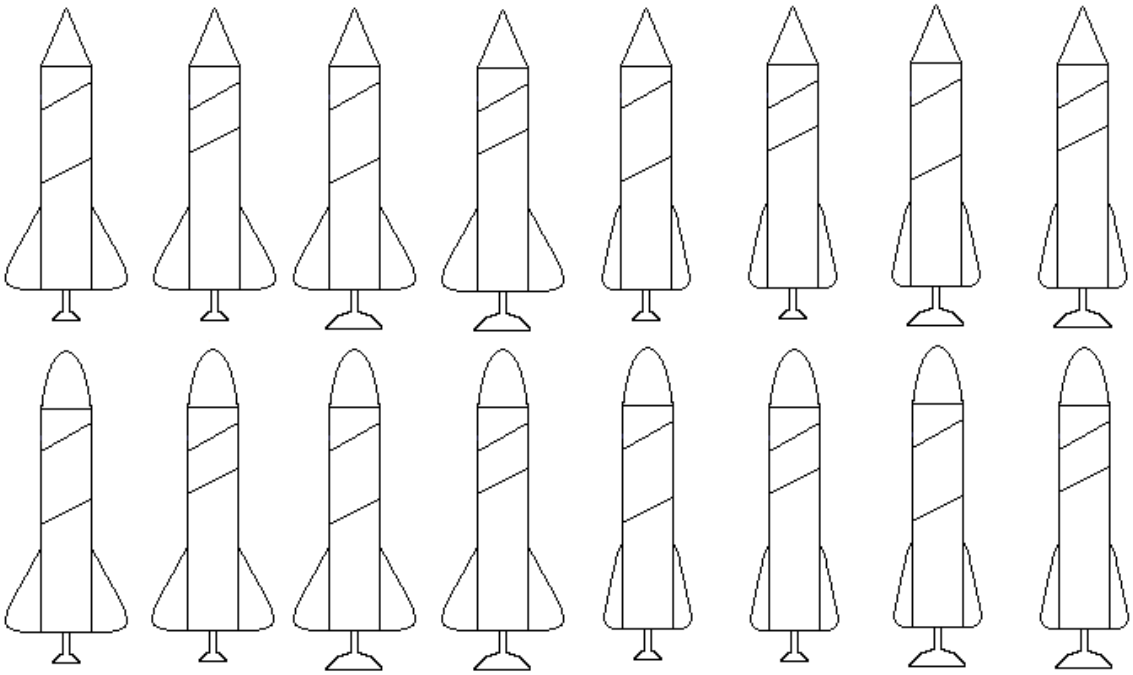

Figure 19. The 16 basic rocket ship stimuli used in Experiments 4-7, composed of binary features on four dimensions: nose cone shape (pointed/rounded), body band size (long/short), wing size (wide/narrow) and booster size (large/small).

Testing trials included a category summary presented on the screen above the testing item and the summary consisted of twelve rocket ships with the category labels underneath, for example see Figure 17. On all trials, participants chose between two on-screen options via key press (w-key response with a "left" sticker on the keyboard for the option on the left side of the screen and p-key response with a "right" sticker for the option on the right). The response options were either the category labels or two different features. After each trial, participants rated their confidence in their answer using a scale from one to nine with one anchored as a rating of 'very unconfident' in their response and nine indicating 'very confident'. 16 classification testing trials individually tested the category assignment of the category summary instances and the generalization instances which were not in the summary, Figure 19.

Following classification testing, this experiment had 46 feature inference testing trials, see Table 3, including the keys tests of generalized and ordinary premise typicality, premise

conclusion similarity and hidden feature attachment. The experiment also included additional tests toward the end, Table 3, that are not central to the key arguments including tests contrasting labels versus non-hidden features and labels versus hidden features, continuous feature tests, etc., as described in Appendix C.

This experiment used three category summary variants, and one of these category summaries was present on the screen while participants were responding on all testing trials, though note that after participants responded the summary was replaced by the response confidence judgment scale. All three category summaries had the twelve instances that matched the abstract category structure, see the first 12 instances in Table 2, six in each category, e.g. Figure 17. On classification testing trials, none of the 12 category instances showed hidden features. However, for all testing trials except the classification, premise diversity and the inclusion fallacy trials, the category summary had hidden feature cut-outs on the two typical and two atypical instances only, e.g. Figure 17. On the premise diversity and the inclusion fallacy trials, see Appendix C for the full description of these testing trials, the category summary contained the hidden features on the typical and atypical instances as well as on two ordinary instances: A1311 and B3133, which were shown to have the typical and atypical hidden feature respectively. So, two instances in each category were shown with the same hidden feature, the typical instance and an ordinary instance for Category A and the atypical instance and an ordinary instance for category B, but these additional hidden features only occurred in the last two testing blocks of the experiment, Table 3.

Participants first read through the on-screen instructions and then completed two practice trials where they were asked to press each of the response buttons to ensure they knew how to respond. They then proceeded through the 62 test trials.

3.2.2.3. Reporting

The full abstract structures for Experiments 4-8 in terms of all testing trials is in Appendix B as well as the average data for each experiment at the level of individual trials. In addition to the key conceptual tests described in detail above, Appendix D discusses the results for additional tests not directly relevant to the key conclusions including: contrasts between category labels and non-hidden and hidden features, classification generalization tests and replications of other less relevant categorical induction effects in the categorical induction paradigm besides the key test of premise typicality.

3.2.3. Results

The classification test results, Figure 20, show a typicality effect as accuracy significantly increased with typicality ($F(2,108) = 12.9$, $p < 0.001$). The typical instance accuracy was significantly higher than ordinary instance accuracy ($t(36) = 6.8$, $p < 0.001$) and ordinary instance accuracy was higher than atypical instance accuracy but not significantly ($t(36) = 1.5$, $p = 0.134$). Overall, participants were reasonably sensitive to the typicality structure of the categories, a necessary prerequisite for a premise typicality effect.

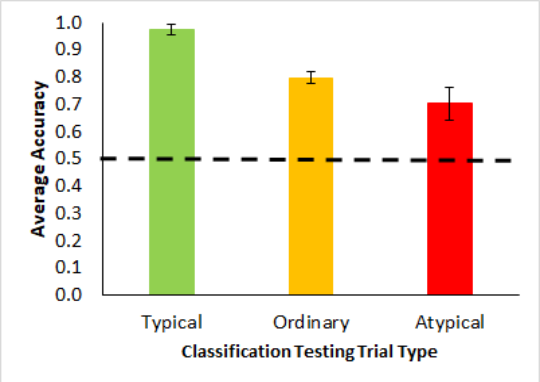

Figure 20. Average accuracy as proportion correct for classification testing trials in Experiment 4, grouped by trial type--typical = green, ordinary = yellow, atypical = red--see Table 2. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Despite sensitivity to the typicality structure, there was no effect of premise typicality, see Figure 21, for the generalized premise typicality tests ($t(36) = 0.1$, $p = 0.891$) or for the ordinary premise typicality tests ($t(36) = 0.9$, $p = 0.378$). Participants showed no preference for responding with the hidden feature attached to the typical instance compared to the hidden feature attached to the atypical instance.

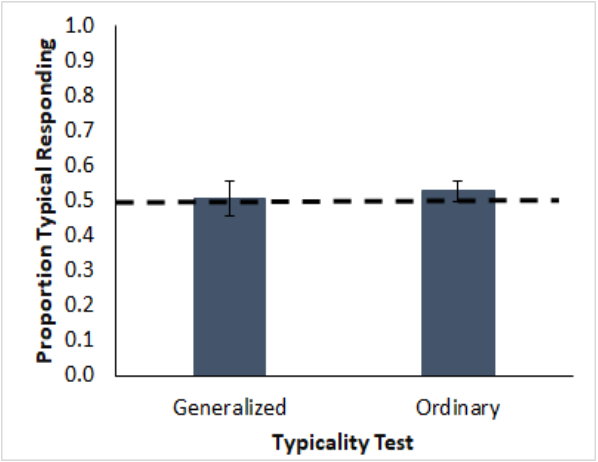

Figure 21. Average proportion of typical hidden feature responding averaged over type of premise typicality trial (Table 3) in Experiment 4. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Clarifying the lack of premise typicality, the hidden feature inference trials, Figure 22 middle bar, showed that the hidden feature attachment to the typical and atypical instances was poor ($t(36) = 0.3$, $p = 0.758$). However, classification testing performance was good ($t(36) = 15.9$, $p < 0.001$) as was the accuracy for the exception feature inference trials ($t(36) = 6.2$, $p < 0.001$), see Figure 22. This suggests that participants were attending to the category labels and non-hidden features but not the hidden features despite their clear presence in the category summary, Figure 17. Lack of attending to the hidden features may have resulted in their not being accurately attached to the typical and atypical instances and this attachment is a necessary prerequisite for a premise typicality effect in terms of having a basis for preferring the typical hidden feature over the atypical.

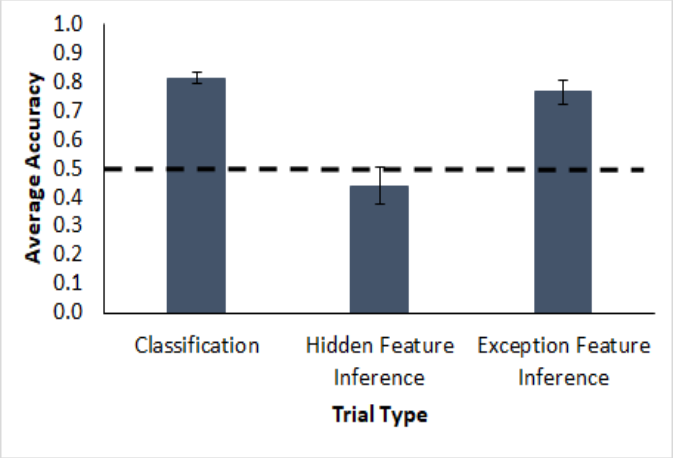

Figure 22. Average accuracy as proportion correct for classification, hidden feature inference and exception feature inference testing trials in Experiment 4, grouped by trial type, see Table 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Despite the lack of premise typicality, as explained by an apparent lack of hidden feature attachment, premise typicality *like* effects occurred in terms of a preference for the typical over the atypical hidden feature. Specifically, premise conclusion similarity trials, Figure 23 left bar, showed significantly more typical hidden feature responding than atypical ($t(36) = 3.6$, $p = 0.001$). Whilst this appears to show premise typicality as a preference for the typical hidden feature, this result is confounded with a difference in similarity in terms of greater similarity to the typical instance than the atypical. Participants were able to respond preferentially with the typical hidden features, when the test item was more similar to the typical than the atypical instance but nonetheless did not show premise typicality when the test item was equal similar, Figure 21. Premise conclusion similarity shows that participants processed and knew something about the hidden features, but strangely this didn't extent to correctly attaching them to the typical and atypical instances on the hidden feature inference trials or to generalizing them to show premise typicality. The blank feature inference trials resulted in another premise typicality *like* effect, Figure 23 right bar, where there was a significant preference for the more typical hidden feature ($t(36) = 3.2$, $p = 0.003$) despite the

presence of only the category label in the testing item. Overall, when the test item was equally similar to the typical and atypical instances, this did not result in a premise typicality effect, Figure 21; however, premise typicality *like* effects occurred, apparently based on the category label alone and on the basis of similarity rather than typicality.

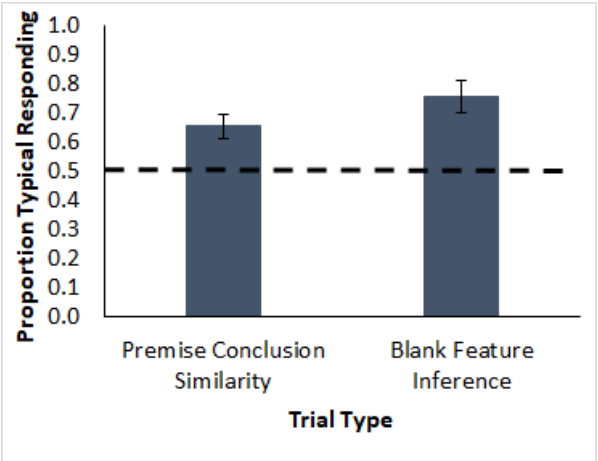

Figure 23. Average proportion typical hidden feature responding for premise conclusion similarity and blank feature inference testing trials in Experiment 4, grouped by trial type, see Table 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

3.2.4. Discussion

In this experiment, participants showed poor performance on attaching the hidden features to the typical and atypical instances despite the correct answers being clearly present in the category summary. This failure to accurately attach the hidden features to the typical and atypical category instances is a plausible reason for the lack of a premise typicality effect despite the fact that they did seem to know something about these features based on the results of the premise conclusion similarity trials. Nevertheless, participants could not apparently generalize these hidden features to new instances based on typicality on the key premise typicality tests. One possible explanation for the lack of hidden feature attachment is that there were many testing trials without clear correct answers, and this may have induced

disengagement with the category summary and guessing. In particular, on the generalized classification trials there were no instances in the category summary that perfectly matched the testing items and this block occurred immediately before the key generalized tests of premise typicality. So, the lack of a sense of giving correct answers based on the lack of clearly matching cases in the category summary may have caused participants to stop engaging with the category summary and start guessing. In order to address this, the next experiment added more trials with clear correct answers.

3.3. Experiment 5

3.3.1. Introduction

Experiment 5 added trials within and between blocks that had a clear correct answer based on the category summary to maintain participants' engagement with the summary. This included adding tests of hidden feature attachment intermixed with the key tests, particularly of premise typicality, as well as three additional classification with hidden features testing blocks. These additions minimized the occurrence of long sequences of trials without clear correct answers. These changes were intended to maintain the participants' attention to the details of the category summary and facilitate better attachment of the hidden features to the typical and atypical instances as a prerequisite for testing premise typicality. Additionally, this experiment added tests of some common categorical induction effects from the classic verbal paradigm, e.g. 'Sparrows have property X Therefore Geese have property X' (example taken from Hayes et al., 2010) at the end of the experiment to ensure that these effects, particularly premise typicality, actually occur in this participant population, see Appendix E.

3.3.2. Materials and Methods

3.3.2.1. Participants

48 Cardiff University students participated for payment or course credit.

3.3.2.2. Materials and Procedure

This experiment differed from the last experiment in terms of adding the four trials that tested hidden feature attachment (Appendix B) to the blocks testing generalized and ordinary premise typicality as well as in the blocks testing premise conclusion similarity and hidden feature inference. Similarly, this experiment added three classification blocks of category instance tests that also had clear correct answers as specific instances in the category summary. The classification blocks occurred before and after the first two premise typicality blocks and after the ambiguous and exception feature inference trial block. To further this, a second block of generalized classification, see Appendix B, occurred with the hidden features present in the category summary and both blocks of generalized classification were moved to the end of testing, thus reducing the number of trials testing an instance without clear correct answers in the category summary before the key tests of premise typicality.

Additionally, this experiment made two minor design changes. First, participants responded via mouse clicking images of the relevant feature/word rather than by pressing buttons on a keyboard as in the prior experiment, and button pressing practice trials were eliminated. Secondly, the experiment changed the trial ordering such that ordinary premise typicality trials occurred before generalized premise typicality trials.

The end of the experiment had 10 classic paradigm categorical induction effect questions using real-world categories that tested premise typicality, conclusion typicality, premise diversity, the inclusion fallacy and premise specificity. These were not based on the perceptual rocket ship stimuli, rather they were representative tests of the classic effect in the standard categorical induction paradigm, taken from Hayes et al. (2010), see Appendix E. The key effect of premise typicality was measured by the difference in likelihood ratings between arguments based on a typical and atypical premise. In the standard effect, the typical premise should be rated a more likely basis for an argument than the atypical premise.

3.3.3. Results

The classification test results, Figure 24, show a typicality effect in that accuracy significantly increased with typicality ($F(2,141) = 18.7$, $p < 0.001$). The typical instance accuracy was significantly higher than ordinary instance accuracy ($t(47) = 7.4$, $p < 0.001$), and ordinary instance accuracy was significantly higher than atypical instance accuracy ($t(47) = 3.9$, $p < 0.001$). This suggests that participants were sensitive to the typicality structure of the categories.

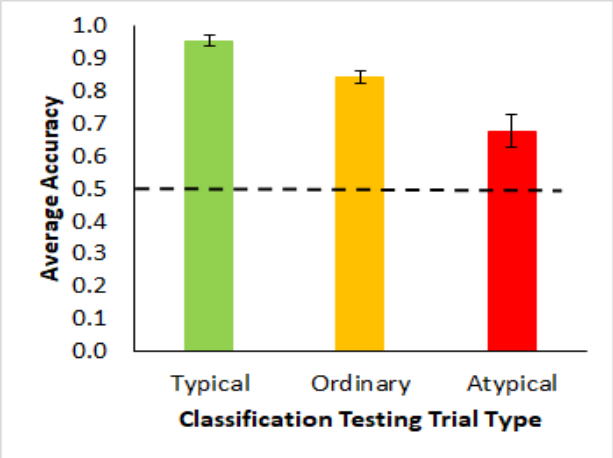

Figure 24. Averaged accuracy as proportion correct for all classification testing trials in Experiment 5, grouped by trial type--typical = green, ordinary = yellow, atypical = red--see Table 2. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Fixing the key lack of hidden feature attachment in the previous experiment, the current hidden feature inference trials, see Figure 25 middle bar, showed good, significant attachment of the hidden features to the typical and atypical instances ($t(47) = 22.4$, $p < 0.001$). Additionally, classification performance across all testing blocks was good ($t(47) = 16.2$, $p < 0.001$) as was exception feature inference performance ($t(47) = 12.8$, $p < 0.001$), see Figure 25. Keeping in mind that the category summary was present on the screen for all testing trials, participants accurately used the summary to classify the instances and attach non-hidden and

hidden features to the instances. Overall this shows engagement with the category summary in terms of good attention to the category label, non-hidden and hidden features, all of which are conceptually necessary for premise typicality.

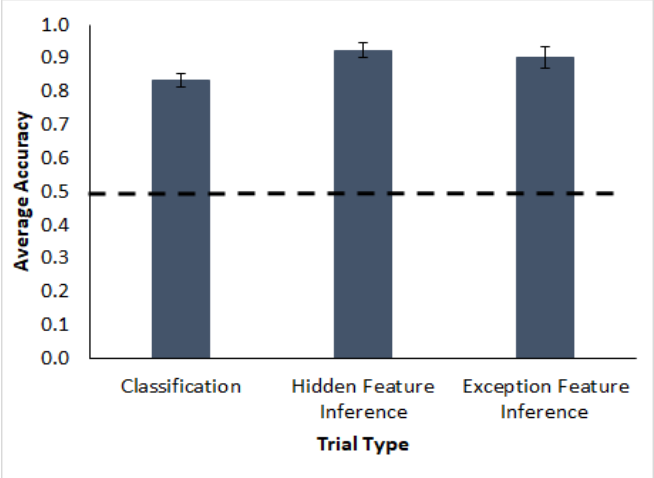

Figure 25. Average accuracy as proportion correct for classification, hidden feature inference and exception feature inference testing trials in Experiment 5, grouped by trial type, see Table 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Despite strong attachment of the hidden features, no premise typicality occurred, see Figure 26, on the generalized premise typicality trials ($t(47) = 1.0$, $p = 0.312$) or on the ordinary

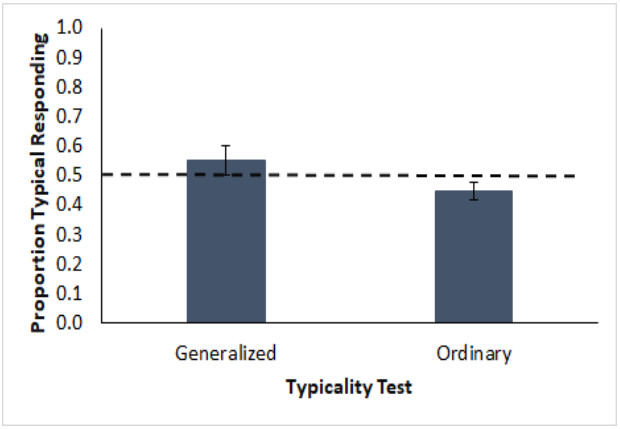

Figure 26. Average proportion of typical hidden feature responding averaged over type of premise typicality trial (Table 3) in Experiment 5. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

premise typicality trials ($t(47) = 1.9$, $p = 0.067$, note: this is in the direction of atypicality not typicality). Participants showed no preference for generalizing the hidden feature attached to the typical instance compared to the atypical instance.

To clarify the strategy used by each participant, error diagrams (similar to those in Chapter 2, Figure 14) show all individual participant responses to classification tests of the summary instances, see Figure 27. White 'dots' represent correct answers on individual trials and black dots represent incorrect answers. Each rectangle is made up of 12 columns which specify the classification trials for all 12 summary category instances (ordered as in Table 2) and four rows which indicate performance on each instance over the four classification testing blocks. The first column of four rectangles labelled, 'Examples', indicates the pattern of responding consistent with a rule on dimensions one through four respectively e.g. a rule on dimension one would be 'a 1 feature on dimension one means the instance belongs to category A, a 3 feature on dimension one means the instance belongs to category B'. Using this rule corresponds to errors on instances A3111 and B1333, see Table 2, and these exceptions to the rule can be seen as vertical black lines of errors in the diagrams. Subsequent columns of rectangles represent participants grouped by performance. The first grouping has participants that responded consistent with one of the four dimensional rules, the second grouping has participants with good overall accuracy, and the third grouping has participants whose responding did not correspond to either of the other groups.

There were two dominant modes of responding, the largest group of participants were responding with high accuracy, 44% of participants, the second largest group were apparently using a dimensional rule, 29% of participants. The other group of participants were using a variety of apparent strategies such as guessing and strategies that resulted in idiosyncratic errors. Thus, of the participants who were engaged with the task, a non-trivial number were

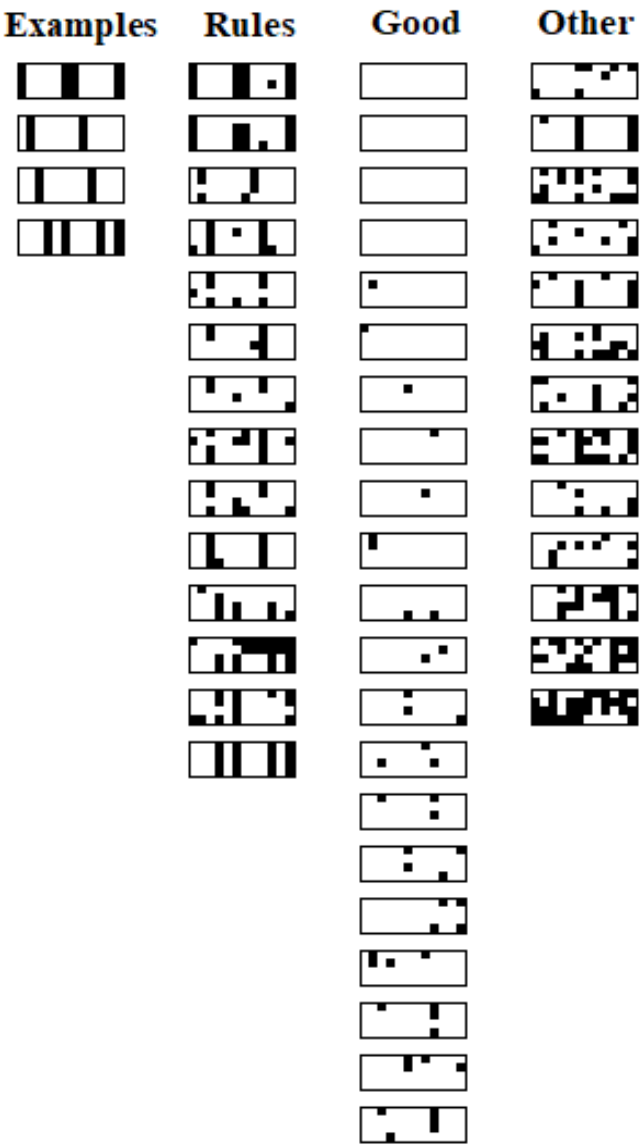

Figure 27. Error diagrams showing performance of each individual participant over classification testing trials for the category summary instances in Experiment 5. Instances are arranged in columns (ordered as in Table 2) and testing blocks are arranged in rows. See main text for details. Black dots = incorrect answers, white 'dots' = correct answers. The 'Examples' grouping shows error patterns corresponding to unidimensional rules in order with a dimension one rule at the top and a dimension four rule at the bottom. The 'rules' grouping has apparent suboptimal dimensional rule users, the 'good' group includes high accuracy performers and the 'other' group has the remaining participants that used various other strategies.

responding consistent with the use of dimensional rules, implying a failure to attend to other features not relevant to their rule and possibly a failure to perceive the typicality structure. If participants did not notice or fully appreciate that the typical instance was more typical than the atypical instance, but rather just classified both typical and atypical instances using a unidimensional rule, then the premise typicality tests may not actually have been a test of premise typicality at all. So, the apparent typicality effects reported above could potentially have been partially due to combining differences in rule use across participants, discussed in detail below.

Classic categorical induction tests based on real categories, Figure 28, showed a significant effect of premise typicality based on a difference in rated argument strength for the typical premise greater than the atypical premise ($t(47) = 4.7$, $p < 0.001$). This replication of the classic paradigm premise typicality effect suggests that the failure to find premise typicality in the perceptual paradigm was not due to a defect in the participant population.

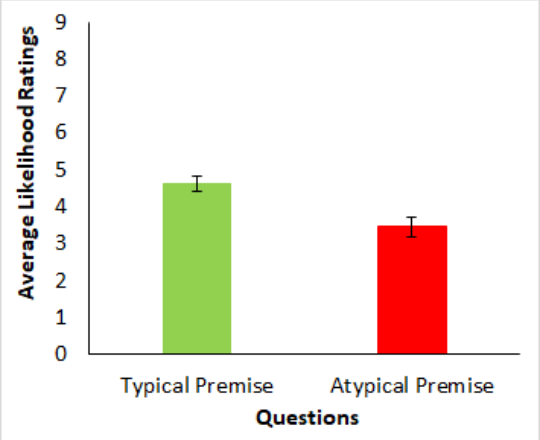

Figure 28. Average argument likelihood ratings for classic paradigm premise typicality testing trials in Experiment 5. Error bars show ±1 standard error.

For the blank feature inference trials, a premise typicality *like* effect occurred in the previous experiment, Figure 23, however this effect did not occur in the current results, Figure 29 right bar; there was not a significant preference for responding with the more typical hidden feature ($t(47) = 1.4$, $p = 0.182$) corresponding to a lack of a premise typicality *like* effect when

only the category label was presented on a testing trial. So, this experiment did not replicate the premise typicality *like* effect from the last experiment (Experiment 4).

Despite the lack of premise typicality, a premise typicality *like* effect occurred in terms of a preference for the typical hidden feature over the atypical hidden feature on the premise conclusion similarity trials, Figure 29 left bar, which resulted in significantly more typical hidden feature responding than atypical ($t(47) = 20.7$, $p < 0.001$). So, a premise typicality *like* effect occurred but likely as a result of a difference in similarity i.e. the test item was more similar to the typical instance than the atypical one. In contrast, when the similarity of the test item to the typical and atypical instances was the same in the premise typicality tests, see Figure 26, participants did not show a preference for the hidden feature associated with the typical instance over the hidden feature associated with the atypical instance.

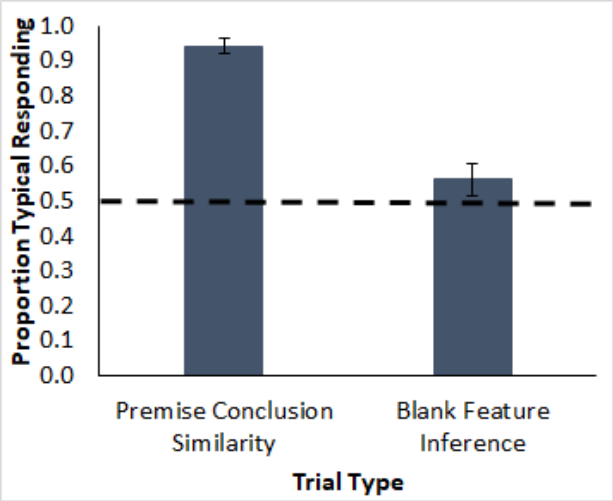

Figure 29. Average proportion typical responding for premise conclusion similarity and blank feature inference testing trials, see Table 3, in Experiment 5, grouped by trial type. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

3.3.4. Discussion

Accuracy was high across the classification testing blocks, but the results also indicated a strong typicality effect as differences in accuracy between the typical, ordinary and atypical instances. Further, participants accurately attached the hidden features to the typical and

atypical instances. However, there was still no effect of premise typicality in the perceptual categorization paradigm despite replicating the classic premise typicality effect based on a written statement such as, 'Sparrows have property X Therefore Geese have property X' (Hayes et al., 2010).

The error diagrams, see Figure 27, show that, despite reasonably high average accuracy, a substantial proportion of participants were not accurately classifying *all* the category instances despite those instances being unambiguously present in the category summary that was always available. Additionally, a nontrivial subset of participants were apparently using suboptimal dimensional rules. This suggests that even mixing in trials with clear correct answers was not enough for the majority of participants to process the category summary sufficiently to do even basic classification. So, while there were "clear" correct answers in the category summary, an obvious reason for these errors was that participants did not receive explicit feedback so the correct answers may not have actually been all that clear for these trials. This suggests including trials which require participants to respond on all dimensions to reduce the effect of ignoring features and dimensions that apparently occurred with unidimensional rule use and also suggests providing explicit feedback.

3.4. Experiment 6

3.4.1. Introduction

The evidence of unidimensional rule use for the classification of instances by a subset of participants in the previous experiment is not consistent with a clear appreciation of the category typicality structure and may even indicate that participants were only actually attending to a single feature dimension of the stimuli. Importantly, rule-based performance can give rise to a pseudo-typicality effect as a result of averaging across participants without individual participants having any understanding of the typicality structure. Stated abstractly, a rule chosen on the basis that a 1 on a dimension belongs to category A and a 3 belongs to

category B will correspond to accurate classification of the typical instances, Table 2. However, each unidimensional rule will cause errors in classifying two ordinary instances, somewhat reducing accuracy for these compared to the typical instances. And two out of the four unidimensional rules will cause additional errors on the atypical instances, reducing accuracy even further compared to the typical instances. Therefore, an apparent typicality effect can occur even if all participants were classifying instances based on unidimensional rules. So again, a key prerequisite for the premise typicality effect is that participants realize that the typical instance is more typical than the atypical instance in order to generalize the typical hidden feature more than the atypical.

Regehr and Brooks (1995) found that the use of a category summary produced single dimensional sorting in categories which is equivalent to unidimensional rule use in a decision-making task. Lassaline and Murphy (1996) found that a way to encourage family resemblance sorting (and therefore encourage an understanding of typicality) was to have participants undergo a task before sorting that facilitated an understanding of the relationship between instances and features. They used perceptual bug stimuli with eight dimensions and either had participants make inferences that emphasized the properties shared between features or make frequency estimates of features which did not encourage this understanding. They found that making feature inferences before a sort encouraged family resemblance sorting compared to those who made frequency judgements or no judgements before the sort. At minimum, this suggests that feature inferences are a good way to get participants to attend to all of the features in the category instances.

Spalding and Ross (1994) showed that a comparison-based learning task presented initially can impact what is learned in a later task by focusing participants on features that were consistent in the earlier comparisons. This has arguably been the case for Experiments 4 and 5 of the current research, where initial classification decision-making comparisons have focused

participants attempting to use rules on whichever of the stimulus dimensions was perceived as more salient as a basis for their rule. A widely found preference for unidimensional rule sorting (Medin, Wattenmaker, & Hampson, 1987) may manifest as a preference to look for unidimensional rules. However, an initial comparison task that forces attention to multiple dimensions, should reduce the preference for unidimensional rules and reduce the number of participants showing a typicality effect without understanding the typicality structure, which should strengthen the basis for premise typicality.

To encourage participants to use all of the feature dimensions in premise typicality decision-making, the present experiment first presented a feedback learning task based on the category summary, including feature inference trials which forced attention to each feature dimension in turn. This additional experience with the category structure was intended to enhance the perception of similarities between instances and features within a category and therefore result in better appreciation of the typicality structure across dimensions, as manifested by reduced dimensional rule use.

3.4.2. Materials and Methods

3.4.2.1. Participants

48 Cardiff University students participated for payment or course credit.

3.4.2.2. Materials and Procedure

The main change in this experiment was the addition of feedback on a series of training trials with the category summary present, before the testing trials. This summary learning task was based on the eight ordinary category instances in Table 2 (excluding the typical and atypical instances for each category) and included eight classification trials and 32 feature inference trials. Each individual feature of the included instances was queried, and participants received feedback for both the classification and feature inference trials. Participants could look at each feedback screen for as long as they wanted and left-clicked the mouse to continue

to the next trial. The eight instances were included as all features of those instances can be unambiguously inferred (when the typical and atypical instances are excluded) and only these eight instances were present in the category summary on the screen during the feedback learning phase. After this the participants completed the same key decision-making tests as in Experiment 5 and also the classic paradigm tests of standard effects including premise typicality questions at the end of the experiment. Finally, the category labels were reduced from two syllables (dreton/rilbar in Experiment 5) to one syllable, 'thab/lork' to make them easier to remember. All other aspects of the experiment were the same as in Experiment 5.

3.4.3. Results

Overall accuracy on the summary learning trials, see Figure 30 first bar, was reasonably good ($t(47) = 11.1$, $p < 0.001$), suggesting that participants were attending reasonably well to all of the feature dimensions and instances. The summary learning classification trials showed good performance, Figure 30 second bar ($t(47) = 9.7$, $p < 0.001$), but all had the same typicality as only the ordinary instances were present in the summary at this point in the experiment. The

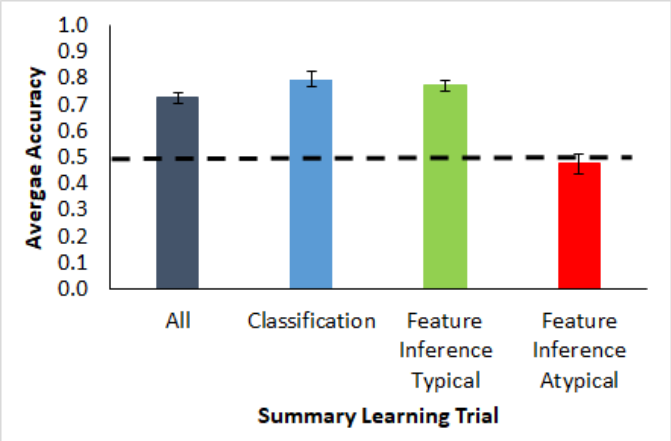

Figure 30. Average accuracy as proportion correct averaged across all learning trials (all data = dark blue), across classification trials (classification = light blue) and averaged across all four blocks of feature inference training trials grouped by trial type (typical = green, atypical = red) for Experiment 6. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

feature inferences were on features that were typical of the category (a '1' feature for category A and a '3' feature for category B) or atypical of the category (a '3' feature for category A and a '1' feature for category B). Participants were significantly more accurate on typical feature inferences than atypical feature inferences ($t(47) = 6.1$, $p < 0.001$) and thus showed an effect of typicality across multiple dimensions, see Figure 30.

The classification test results for the ordinary, typical and atypical instances were without feedback and show, Figure 31, a typicality effect as accuracy increased with typicality ($F(2,141) = 7.9$, $p = 0.001$). There was a non-significant trend for typical instance accuracy to be higher than ordinary instance accuracy ($t(47) = 1.2$, $p = 0.246$), and ordinary instance accuracy was significantly higher than atypical instance accuracy ($t(47) = 4.1$, $p < 0.001$). The feature inference feedback trials and the classification testing trials show sensitivity to the typicality structure of the category across dimensions, fixing the apparent issue in the previous experiment that some participants were seemingly attending to only one dimension as indicated by the error diagrams, see Figure 27.

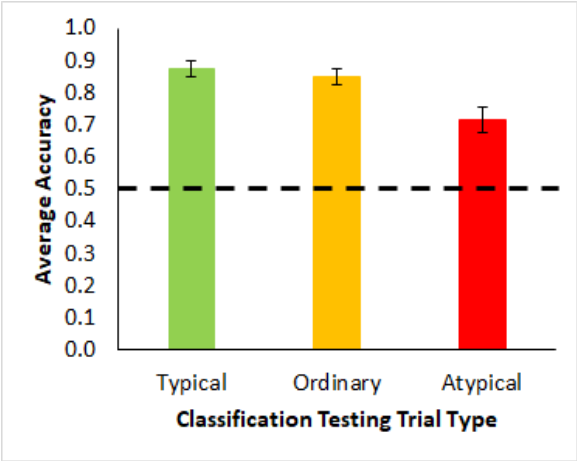

Figure 31. Averaged accuracy as proportion correct averaged across all blocks of classification testing trials, see Table 2, for Experiment 6, grouped by trial type--typical = green, ordinary = yellow, atypical = red. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Confirming this reduction in dimensional rule use, the error diagrams (explained on p. 81) in the current experiment showed only 6% of participants using rules, see Figure 32. This suggests that the typicality results were not due to apparent rule users giving typicality like responses without sensitivity to the typicality structure. So, a potential reason for the lack of premise typicality in the prior experiment was eliminated in this experiment.

The hidden feature inference trials, see Figure 33 middle bar, showed good attachment of the hidden features to the typical and atypical instances ($t(47) = 13.0$, $p < 0.001$), maintaining the high levels of attachment demonstrated in Experiment 5. Additionally, classification performance, see Figure 33 left bar, across all testing blocks was good ($t(47) = 14.3$, $p < 0.001$) as was the exception feature inference ($t(47) = 7.8$, $p < 0.001$), see Figure 33 right bar. Keeping in mind that the category summary was present on the screen for all testing trials, participants showed high levels of engagement with the category summary particularly as high accuracy in attaching the hidden features to the typical and atypical instances. Again, this is a key prerequisite for a premise typicality effect.

Despite good attachment of the hidden features to the typical and atypical instances and sensitivity to the typicality structure of the category, no premise typicality occurred, see Figure 34, on generalized premise typicality trials ($t(47) = 0.4$, $p = 0.681$) or ordinary premise typicality trials ($t(47) = 0$, $p = 1$, note that the average proportion was literally 0.50). Participants showed no preference for generalizing the hidden feature from the typical instance compared to the atypical instance when similarity of the test instance to the typical and atypical instances was the same.

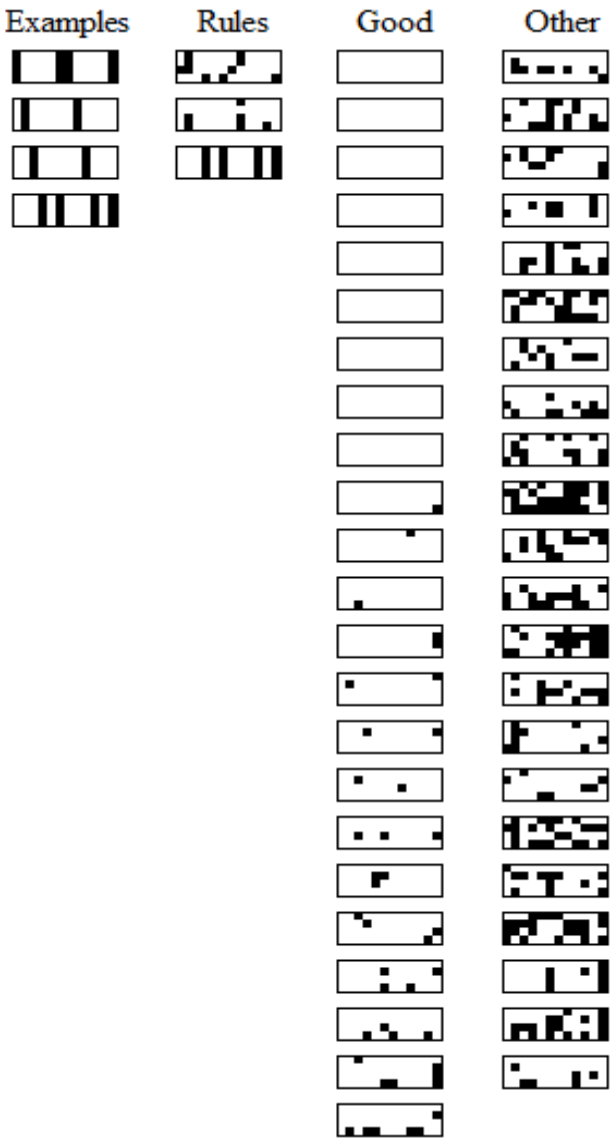

Figure 32. Error diagrams showing performance of each individual participant over classification testing trials for the category summary instances in Experiment 6. Instances are arranged in columns (ordered as in Table 2) and testing blocks are arranged in rows. See main text for details. Black dots = incorrect answers, white 'dots' = correct answers. Error patterns in the 'Examples' grouping correspond to unidimensional rules, shown in order with a dimension one rule at the top and dimension four rule at the bottom. The 'rules' grouping has apparent suboptimal dimensional rule users, the 'good' group includes high accuracy performers and the 'other' group has the remaining participants that used various other strategies.

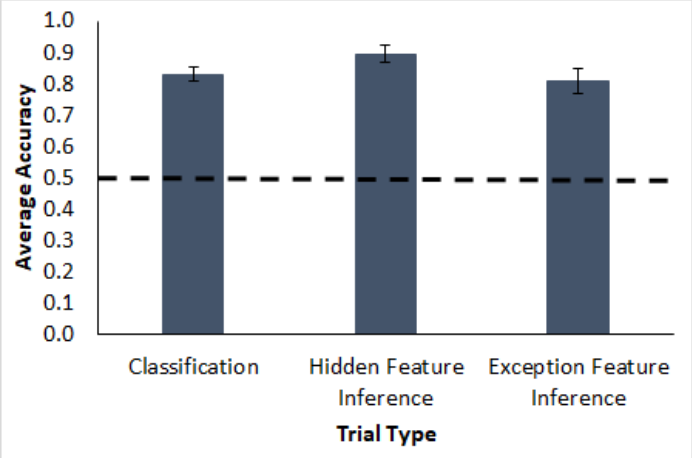

Figure 33. Average accuracy as proportion correct for the classification, hidden feature inference and exception feature inference testing trials, see Table 3, in Experiment 6. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

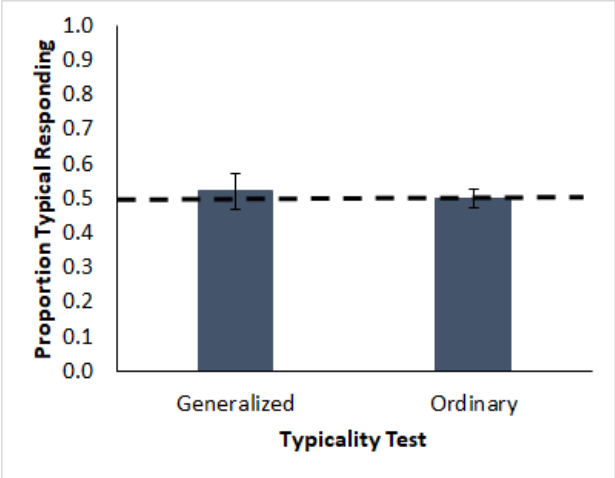

Figure 34. Average proportion of typical hidden feature responding averaged over type of premise typicality trial (Table 3) in Experiment 6. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

The classic paradigm tests of premise typicality, Figure 35, showed a significant effect of premise typicality based on a difference in rated argument strength for the typical premise greater than the atypical premise ($t(47) = 2.5$, $p = 0.014$). This replicates the classic paradigm effect, further confirming the conclusion that the failure to find this effect in the perceptual paradigm was not due to a defect in the participant population.

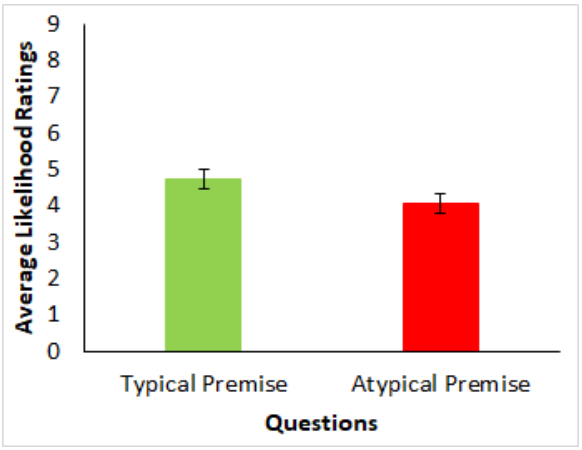

Figure 35. Averaged argument likelihood ratings for classic paradigm premise typicality testing trials in Experiment 6. Error bars show ±1 standard error.

Consistent with the lack of premise typicality, the blank feature inference trials, Figure 36 right bar, showed no premise typicality *like* effect ($t(47) = 1.1$, $p = 0.280$). However, a premise typicality *like* effect occurred on the premise conclusion similarity trials, Figure 36 left bar, which had significantly more typical hidden feature responding than atypical ($t(47) = 5.6$, $p < 0.001$). As in the previous experiment, this shows a premise typicality *like* effect in which similarity was confounded with typicality. However, when the test item was equally similar to the typical and atypical instances, participants showed no premise typicality effect, Figure 34.

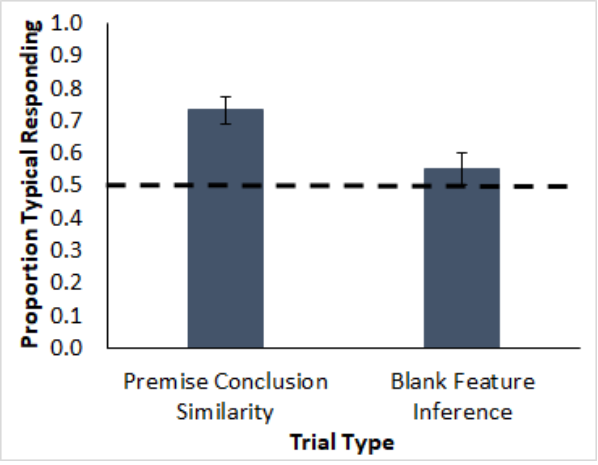

Figure 36. Average proportion typical responding for premise conclusion similarity and blank feature inference testing trials, see Table 3, grouped by trial type in Experiment 6. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

3.4.4. Discussion

In this experiment accuracy was high across the summary learning task and across multiple classification testing blocks. Additionally, high accuracy on the typical testing items compared to the atypical items showed an effect of typicality validated by the relative absence of apparent dimensional rule users. Further, participants accurately attached the hidden features to the typical and atypical instances. However, there was no effect of premise typicality despite an effect being found in the classic paradigm categorical induction task.

Regehr and Brooks (1995) found that the use of a category summary encourages unidimensional rule use, and Medin et al. (1987) reported a preference among participants to use unidimensional rules in sorting based decision-making tasks. However, as Lassaline and Murphy (1996) suggested, this experiment included a task before decision-making which encouraged a focus on more feature dimensions and unidimensional rule use declined. Similarly, the current results are consistent with those of Spalding and Ross (1994) which suggested that an initial comparison-based learning task can influence later performance as the introduction of the summary learning task in this experiment reduced the number of rule users. Overall, the inclusion of the summary learning task had the desired effects in terms of encouraging participants to attend to all feature dimensions and reducing unidimensional rule use; however, the current findings suggest that rule use was not the basis for the lack of the premise typicality effect.

3.5. General Discussion

Experiment 4 employed a basic summary decision-making task to test premise typicality. Whilst participants were able to correctly classify category instances and were apparently sensitive to the typicality of the structure and to similarity effects, participants did not correctly attach the hidden features to the typical and atypical instances and did not show premise typicality. Experiment 5 remedied this with the inclusion of more intermixed trials that

had clear correct answers in the category summary to motivate participants to continue using the summary even after tests on items with no perfectly matching correct answers in the summary. This manipulation was successful, and participants on average attached the hidden features to the typical and atypical instances accurately. However, they were not able to generalize this to use the typical hidden feature more than the atypical hidden feature and did not show an effect of premise typicality. However, the error diagrams indicated quite a few participants who apparently used unidimensional rules. As previously described (pp. 85-86), participants using unidimensional rules to make classification decisions can produce pseudo-typicality: a pattern of responding that looks like a typicality effect but in the absence of any appreciation of typicality. To reduce dimensional rule use, Experiment 6 used a feedback learning task based on the category summary before the decision-making task to ensure that participants focused on all non-hidden feature dimensions, by querying features on each dimension before the critical premise typicality trials. This is in contrast to participants potentially attending to only one dimension when using a unidimensional rule. Again, participants correctly classified the category summary instances and showed typicality effects with very little evidence of dimensional rule use. Further, participants were accurate in attaching the hidden features to the typical and atypical instances, however there was still no evidence of premise typicality.

The error diagrams for Experiment 5 suggested that roughly a third of participants were using rules, a third were using a strategy such as overall similarity that allows perfect performance, and some were responding randomly. This is consistent with the findings of Little and McDaniel (2015) who found that participants differentially used rules or memorization strategies on the same task, and in particular, some used perceptual similarity for ambiguous items. This shows individual differences in categorization strategy across participants. Pothos (2005) argued that rules and similarity are on a continuum in which using rules equates to using

similarity on the basis of a restricted set of features. The participants in the current experiments had the category summary available when responding and therefore had at least the potential to compare a testing instance to other instances and to multiple features in the category summary. The extent to which this occurs can be argued to depend on the level of experience participants have with the category summary and all of its components. Those that had less experience (potentially due to limited desire to attend) may have used a single dimension-based strategy and those who were more engaged, that is, more experienced (due to higher levels of attending) may have used more features. This is consistent with the groupings of participants in the error diagrams and suggests that with an increase in experience participants can be encouraged to reduce their attention to single dimensions and reduce attempted dimensional rule use as was seen in Experiment 6.

Further, Johansen and Palmeri (2002) found that the strategy for classifying instances can change over the course of an experiment from dimensional rule use initially to responding based on multiple dimensions with experience. Therefore, experience could be what allows participants to stop using unidimensional rules and attend to all dimensions. Participants who lack the opportunity (or motivation) to gain enough experience of the category structure may not make this change. The inclusion of summary learning trials with feedback resulting in less rule use in Experiment 6 also supports this idea and suggests that a more in-depth learning task might be sufficient to produce a premise typicality effect, as evaluated in the next chapter.

To summarize, there are certain minimal performance requirements for a perceptual categorization task testing premise typicality that need to be met, as the absence of any of these prerequisites constitutes a plausible explanation for the lack of premise typicality. I argue that there are three such prerequisites and discuss a fourth, the lack of internal category representation, which may be the explanation for the lack of premise typicality in Experiments 4-6. First, participants must have the ability to correctly categorize the category instances as

indicated by high accuracy on the classification testing trials. Second, participants must demonstrate understanding of the typicality gradient within the category structure as indicated by, for example, higher accuracy for the typical instances than the atypical instances. This ensures that when participants generalize a feature, they at least have the potential to understand that the typical instance is better for such generalization than the atypical instance. Thirdly, they must correctly attach the hidden features to both the typical and atypical instances as indicated by high accuracy on the hidden feature inference trials. Without this, participants may be aware that one instance is more typical than another, and so is potentially better for the generalization of hidden features, but not know what that hidden feature is.

In Experiments 5 and 6 all three prerequisites were met and yet premise typicality did not occur. For this reason, I pose a fourth prerequisite as a possible requirement: a strong *internal* mental representation of the categories. This is in contrast to the category summary used in the current experiments which is an *external* representation of the categories. If the internal representation of the categories is developed by a full learning task, this has the potential to strengthen the participants' natural understanding of a category's typicality structure, allowing them to infer hidden features consistent with that structure. This aligns closely to the well-known demonstrations of premise typicality in the classic paradigm version of categorical induction as in that domain the categories used are real-world, well known categories such as bird or mammal with representations that are definitely internal.

Chapter Four - Premise Typicality as Feature Inference Decision-Making following Classification Learning

4.1. General Introduction

Perceptual learning is a basic cognitive skill required for everyday functioning in the world. Many studies (e.g. Ashby & Gott, 1988; McKinley & Nosofsky, 1995; Medin & Schaffer, 1978; Shephard et al., 1961; Yamauchi & Markman, 1998; Yamauchi et al., 2002; etc.) have used carefully designed category structures in the perceptual learning paradigm to draw conclusions about category representation. For example, Medin and Schaffer (1978) concluded that people represent categories by storing instances/exemplars and Shepard et al. (1961) concluded that people use rules, as discussed in Chapter 2. The present experiments were based on a variant of the family resemblance category structure (Rosch & Mervis, 1975) used in Experiments 4-6 (Chapter 3) where the instances varied in typicality with a highly typical prototype, four ordinary instances each with one prototype inconsistent feature, and an atypical instance with two prototype inconsistent features.

Experiments 4-6 (Chapter 3) evaluated premise typicality via decision-making and via a learning task based on the external summary representation of family resemblance categories (Figure 17 and Table 2 in Chapter 3). Premise typicality did not occur in any of these tasks, possibly due to the use of the external representation and the absence of a fully internalized representation of the categories, analogous to the internal category representations in the classic categorical induction paradigm, e.g. for categories such as robins and sparrows, etc.

Given that premise typicality is measured via hidden feature inferences, one possible learning approach to producing a fully internalized representation is feature inference learning. However, Sweller and Hayes (2010) showed that feature inference learning of both typical and atypical features is difficult and unreported data I have collected showed that almost no participants were able to learn this category structure by feature inference of instance features.

Further, Jee and Wiley (2014) found that classification learning allowed a better understanding of the typical and atypical features within a category than feature inference learning. This suggests that classification learning is a reasonable approach to generating a sense of instance typicality.

Chapter 3 (pp. 96-97) identified four prerequisites, the absence of any of which could provide a reason for the observed absence of premise typicality. As such, these prerequisites needed to be met for the current tests to be a legitimate assessment of the presence or absence of a premise typicality effect in this paradigm. The first is accurate classification testing performance. In the summary decision-making task, failing to accurately classify the instances in the summary likely indicated a lack of engagement with the summary as the task wasn't particularly difficult. Participants could satisfy the easy requirement of accurately classifying the instances in the summary without internally representing the categories. Without the category summary the task presumably becomes harder; however, accurate classification performance would imply a strong internal category representation which provides a basis for appreciating the typicality structure and therefore a basis for preferring the typical hidden feature in the key premise typicality test.

The second prerequisite is appreciation of the typicality structure of the categories as indicated by a typicality effect, for example, higher accuracy for the typical category instance than for the atypical instance. This appreciation of the typicality structure in the category potentially allows a realization that some instances are a better basis for generalizing hidden features than others. If participants don't perceive that some instances are more typical than others then a preference for one hidden feature over the other cannot be based on typicality, a defining precondition for evaluating premise typicality. On the one hand, the external category summary seems particularly conducive to noticing which features are typical, but on the other,

forgetting and interference between category instances internally in memory may be necessary to foster a sense of typicality.

The third prerequisite is the accurate attachment of the hidden features to their respective typical and atypical instances. If participants cannot accurately identify the hidden features associated with the typical and atypical instances, then they have no basis in typicality for preferring the typical hidden feature over the atypical on a test of premise typicality. When the categories are externally represented in a summary, then a failure to accurately attach a hidden feature to a typical or atypical instance is likely due to a lack of engagement as the feature is visually associated with only one instance. Attachment of the hidden features is more challenging to represent internally for some types of representation but is still a necessary methodological prerequisite before asking about a preference for the typical hidden feature over the atypical on a test of premise typicality.

After the failure to find premise typicality in decision-making tasks based on a category summary, these three prerequisites converge to one potentially fundamental prerequisite: the internal representation of the categories. Specifically, a category may need to be learned and mentally represented internally, as are the categories in the classic categorical induction paradigm, e.g. robins, to fully appreciate the typicality structure they have. The perceptual category learning paradigm has been widely shown to generate fully internalized category representations that produce category typicality effects (Light et al., 1979; Lin et al., 1990; McCloskey & Glucksberg, 1978; Medin & Schaffer, 1978; Nosofsky, 1988; Rosch & Mervis, 1975; Rosch et al., 1976; Rothbart & Lewis, 1988; Spalding & Murphy, 1999; etc.) and as such, perceptual learning tasks should produce analogues of premise typicality. The following two experiments were designed to satisfy the fourth prerequisite in terms of producing a fully internalized category representation while continuing to satisfy the first three prerequisites.

4.2. Experiment 7

4.2.1. Introduction

The purpose of this experiment was to use feedback learning of category instances across a series of trials to produce a fully internalized representation of the family resemblance category structure, see Table 2 in Chapter 3, before the key test of premise typicality. Premise typicality was tested by querying hidden features and is described in detail in Chapter 3 (p. 66). To highlight the typicality structure in the categories, this experiment included a phased introduction of the typical and atypical instances. The ordinary category instances were introduced first in early learning blocks, followed by blocks of the typical and atypical instances with their hidden features. This was intended to sharply distinguish the typical and atypical instances from the ordinary instances and highlight the attachment of the hidden features to those instances.

4.2.2. Materials and Methods

4.2.2.1. Participants

48 Cardiff University students participated for course credit or payment.

4.2.2.2. Materials and Procedure

The learning phase in this experiment eliminated the category summary from Experiment 6 and instead presented a series of classification trials with feedback. The learning task introduced instances in the category structure in a phased way with just the ordinary category instances (excluding the typical and atypical instances) in the first 20 blocks (160 trials) followed by 20 blocks (80 trials) of just the typical and atypical instances with their hidden features.

The testing phase added an extra block each of the ordinary and generalized premise typicality trials and all the premise typicality testing blocks also included tests of hidden feature

attachment for the typical and atypical instances. Additionally, two blocks of classification testing trials were removed as participants no longer needed the additional trial blocks that fostered interaction with the category summary as the category summary was eliminated. All other aspects of the experiment were the same as in Experiment 6 (though see Appendix B for minor changes to one block of generalized classification trials).

4.2.3. Results

Average accuracy in the first 20 blocks of the learning phase, the orange line Figure 37, was not particularly good at 68% correct. As such, a post data collection criterion was imposed of greater than 75% correct over the last 4 blocks of ordinary instance learning, i.e. over blocks 17-20, blue line Figure 37, and the resulting average performance in those blocks was 94% correct. 15 participants met this criterion; so, about a third of participants learned the task well. Only these criterion data are considered in subsequent analyses with the exception of the tests of premise typicality and the error diagrams.

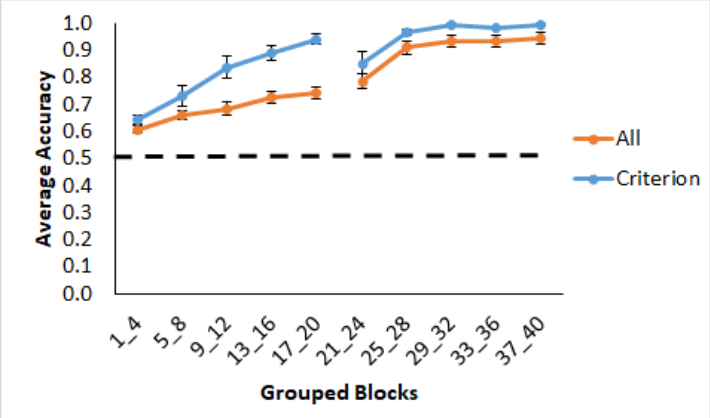

Figure 37. Average accuracy as proportion correct across all learning blocks with all data as orange lines and criterion data (75% or greater correct in blocks 17-20) as blue lines for Experiment 7. Blocks 1-20 were ordinary instance classification learning trials and blocks 21-40 were typical/atypical instance classification learning trials. The two learning phase learning curves are separated by a break in the lines. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

The classification learning results, blocks 1-40 in Figure 38, show an initial typicality effect as significantly higher accuracy on the typical than the atypical instances in blocks 21-24 ($t(14) = 3.2$, $p = 0.006$). While ordinary instance accuracy was significantly lower than both the typical instances, ($t(14) = 8.8$, $p < 0.001$) and the atypical instances, ($t(14) = 5.7$, $p < 0.001$), ordinary instances were trained separately in earlier blocks (1-20). Accuracy remained high in the first classification testing block, CL1 on the far right in Figure 38, but the drop in accuracy for the atypical instances in CL2 left overall accuracy on the typical instances significantly higher than the atypical instances, ($t(14) = 4.1$, $p = 0.001$). Overall the phased introduction of the typical and atypical instances (blocks 21-40) after the ordinary instances (blocks 1-20) apparently made the atypical instances especially easy to learn, but they were still learned more slowly than the typical instances, consistent with a typicality effect.

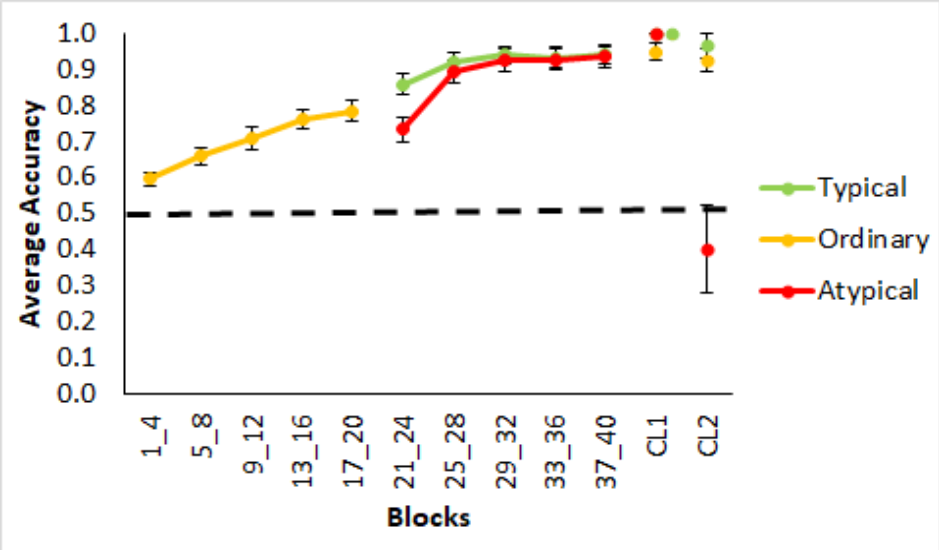

Figure 38. Averaged accuracy as proportion correct for classification learning in Experiment 7, grouped by trial type--typical = green, ordinary = yellow, atypical = red--see Table 2 in Chapter 3. Dots on the far right indicate average accuracy in the classification testing blocks, CL1 and CL2, for different levels of typicality. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Overall classification testing performance, combined across all typicality instances, was good, Figure 39 left bar ($t(14) = 26.2$, $p < 0.001$), indicating the persistence of learning from the training phase. However, accuracy for the hidden feature inference trials was poor, Figure 39 middle bar ($t(14) = 0.9$, $p = 0.384$), suggesting participants did not successfully attach the typical and atypical hidden features to the typical and atypical category instances. Exception feature inference was also poor, Figure 39 right bar ($t(14) = 0.4$, $p = 0.709$). So, while participants knew the category structures well enough to classify the instances, including those with hidden features visible, there was little evidence that they could make accurate feature inferences, especially of the critical hidden features, a necessary prerequisite for premise typicality.

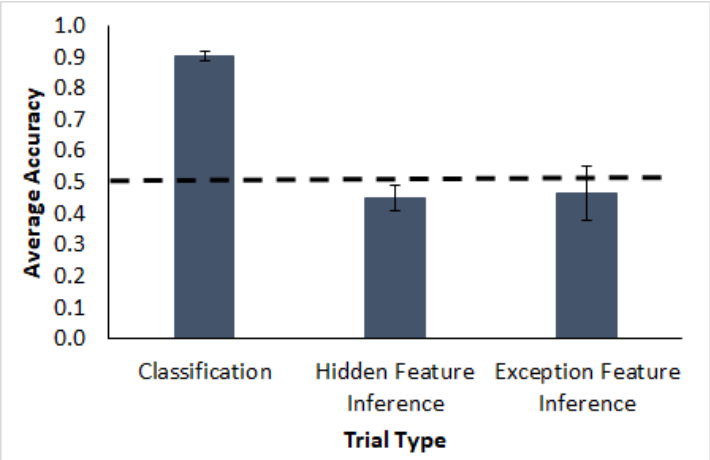

Figure 39. Average accuracy as proportion correct for the classification, hidden feature inference and exception feature inference testing trials in Experiment 7, grouped by trial type, see Table 3 in Chapter 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

Consistent with the poor attachment of the hidden features, there was no effect of premise typicality for participants who met the criterion, Figure 40 blue bars, on the generalized premise typicality tests ($t(14) = 0.4$, $p = 0.683$) or the ordinary premise typicality tests ($t(14) = 0.7$, $p = 0.510$). There was also no effect of premise typicality for all participants including those who did not meet the learning criterion, Figure 40 orange bars, on the generalized premise

typicality tests, ($t(47) = 0.3$, $p = 0.772$), or the ordinary premise typicality tests ($t(47) = 0.1$, $p = 0.930$).

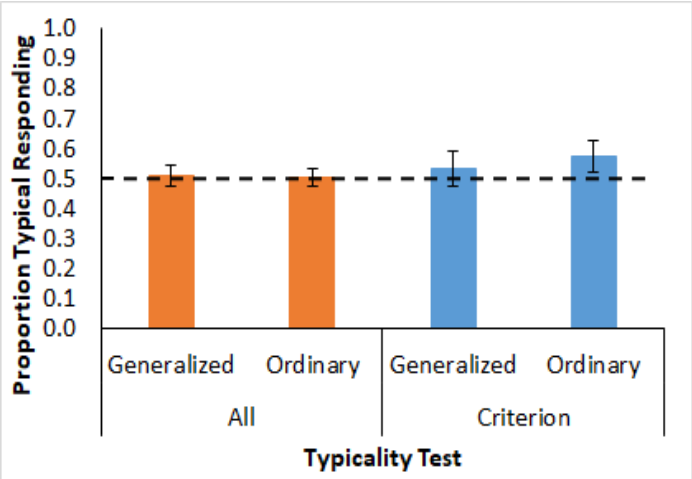

Figure 40. Proportion of typical hidden feature responding averaged over both blocks of each type of premise typicality testing trial (Chapter 3, Table 3) in Experiment 7. Orange bars = all data, blue bars = criterion data. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

As in the decision-making experiments in Chapter 3, error diagrams for all participants, Figure 41, are helpful in identifying participants apparently using dimensional rules as this is potentially a marker for task difficulty and a tendency to attend to only a subset of dimensions. The error diagrams here have a somewhat different format than those in the previous chapter in that rows and columns are swapped (see Chapter 3, p. 81); specifically, each column here has the results of a single training block, and each row indicates accuracy on a particular instance. So here, the errors associated with being exceptions to a dimensional rule correspond to horizontal black lines. In addition, the present diagrams had the added high-level grouping, 'Good HFs' containing participants who were reasonably accurate on the typical and atypical instances with hidden features in blocks 21-40 but not accurate on the ordinary instances in blocks 1-20. The good performance on both the typical and atypical instances in the absence of good performance on the ordinary instances suggests that good performance didn't arise out

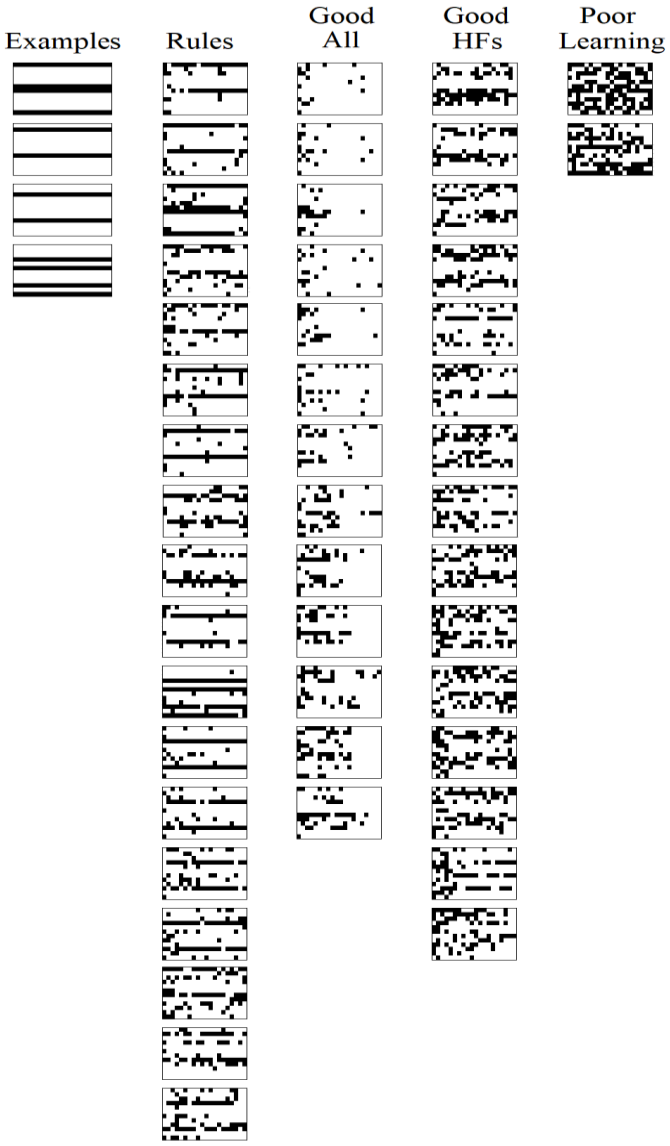

Figure 41. Error diagrams showing performance of each individual participant over all classification learning trials in blocks 1-40 of Experiment 7. Instances are arranged in rows (ordered as in Table 2) and learning blocks are arranged in columns. Black dots = incorrect answers, white 'dots' = correct answers. The 'Examples' grouping shows error patterns corresponding to unidimensional rules in order with a dimension one rule at the top and a dimension four rule at the bottom. The 'Rules' grouping has apparent dimensional rule users, the 'Good All' group includes performers with high accuracy on all trials, the 'Good HF' group includes performers with high accuracy on the typical and atypical instances only and the 'Poor Learning' group has low accuracy performers.

of an appreciation of the typicality structure. Rather, this good performance may have been based on a dimension two or three rule which would have allowed perfect performance in blocks 21-40 as these only included the typical and atypical instances *not* the ordinary instances. This allows the rules on dimensions two and three to be perfectly diagnostic. Related to this, suboptimal, dimensional rule users were the largest group in Figure 41 with 38% of participants, while the Good HFs learning group were second largest at 31%. Taken together, this suggests that many participants were using dimensional rules especially in the typical/atypical phase of learning. As discussed in Chapter 3 (pp. 85-86), averaging across participants using different dimensional rules can look like a typicality effect even though individual participants may not have had any appreciation of the typicality structure of the categories. At minimum, the number of dimensional rule users is symptomatic of task difficulty and a tendency to not attend to most of the stimulus dimensions which is problematic as the differences in category instance typicality are specified in terms of most or all of the dimensions.

The classic paradigm version of premise typicality, Figure 42, resulted in a significant premise typicality effect with greater likelihood ratings over all participants for the argument including the typical premise compared to the argument with the atypical premise, ($t(47) = 2.0$, $p = 0.05$). However the effect was not significant for the learning criterion participants, ($t(14) = 0.5$, $p = 0.655$), likely due to the reduction in power from the small number of participants.

In this experiment there were no premise typicality effects, likely due to the lack of hidden feature attachment to the typical and atypical instances, and similarly there were no premise typicality *like* effects on the blank feature inference testing trials, Figure 43 right bar. There was no significant preference for responding with the typical hidden feature over the atypical hidden feature, ($t(14) = 0.8$, $p = 0.433$). The premise conclusion similarity trials, Figure 43 left bar, produced a trend in terms of a preference for the typical hidden feature over the

atypical, but unlike the prior results in Experiments 4, 5 and 6 (Chapter 3), this was not significant, ($t(14) = 1.1$, $p = 0.290$).

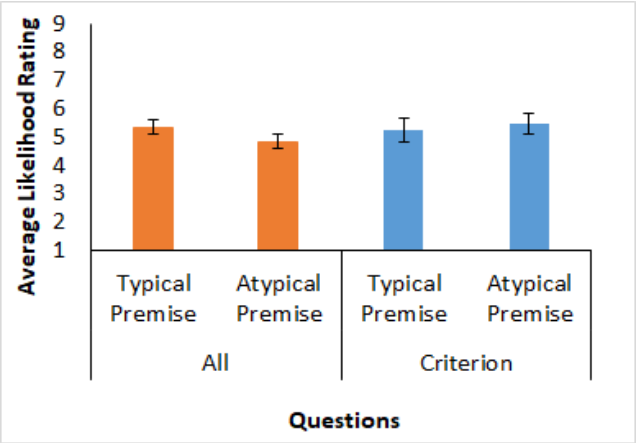

Figure 42. Average argument likelihood ratings for the classic paradigm premise typicality testing trials in Experiment 7. Orange bars = all data, blue bars = criterion data. Error bars show ±1 standard error.

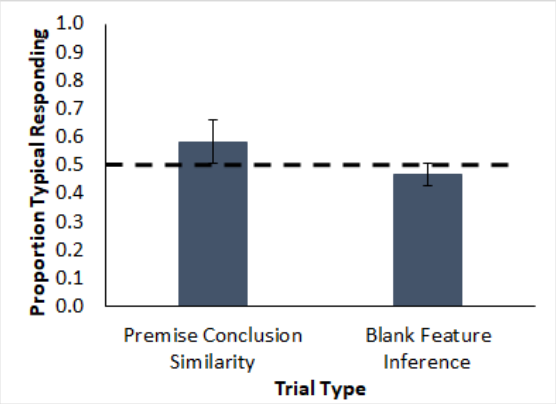

Figure 43. Average proportion typical hidden feature responding for premise conclusion similarity and blank feature inference testing trials in Experiment 7, grouped by trial type, see Table 3 in Chapter 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

4.2.4. Discussion

This experiment implemented a full learning task to encourage the formation of a fully internalized category representation (prerequisite 4), before the key tests of premise typicality. A criterion ensured that only those who learned the task reasonably well were included in the

testing trial analysis (prerequisite 1, the accurate classification of category instances). Prerequisite 2 was apparently satisfied in terms of a typicality effect as accuracy on the typical instances was higher than on the atypical instances. However, this may not have been based on an actual appreciation of the typicality structure of the category as participants were more accurate on the atypical instances during learning than the ordinary instances, despite having more prior training on the ordinary instances. This suggests that participants used dimensional rules during the second phase of learning as, for example, the rule, 'an instance that has a 1 feature on dimension 2 is a thab and a 3 feature on dimension 2 is a lork,' allows perfect performance and therefore higher asymptotic accuracy on the atypical instances than on the ordinary instances as an artefact of the phased design. Moreover, attempts to use dimensional rules in phase 2 are consistent with the observed typicality effect in the learning criterion participants in that early attempts to use rules on dimension one or four by some participants would accurately classify the typical instances but not the atypical, while a rule on dimension two or three would accurately classify both. This could give an initial learning decrement in performance on the atypical instances. The existence of the "Good HF" group in the error diagrams also supports the occurrence of this strategy in terms of high accuracy on the typical and atypical instances despite low accuracy on the ordinary instances. The use of such a dimensional rule in phase 2 does not correspond to participants appreciating the typicality structure of the categories and does not provide a basis to learn the attachment of the hidden features to the typical and atypical instances. And this is consistent with the key limitation of this experiment in terms of the failure to satisfy prerequisite 3 as indicated by the observed lack of hidden feature attachment in the results. As previously stated, premise typicality cannot be reasonably assessed unless participants accurately attached the hidden features to the typical and atypical instances, otherwise they have no basis for preferring one hidden feature over the other on the key tests of premise typicality. To remedy this, the learning phase in the next

experiment still introduced the hidden features in a phased way but also continued to train the ordinary trials into the second learning phase so as to discourage participants from trying to use a separate strategy for classifying the typical and atypical instances.

The error diagrams suggest that rule use was also a prevalent strategy for learning the ordinary instances in the first learning phase, yet a further symptom of a tendency to not attend to all of the stimulus dimensions. A plausible reason that participants use rules is that unidimensional rules are easy to apply in what is a relatively difficult learning task and use of these rules produces better than chance performance. This difficulty of learning is indicated by only 31% of participants reaching the learning criterion of greater than 75% accuracy over the final four blocks of ordinary instance learning. There are several potential ways to make this learning task easier. One is to improve the discriminability and verbalizability of the stimuli, as described in detail in Chapter 2 (pp. 39-40). Kurtz et al. (2013) found that the verbalizability of the stimuli features and dimensions used in a perceptual learning task impacted the ease of learning. For example, when a feature dimension consisted of difficult to describe patterns instead of easy to name solid colours, performance on an Exclusive-Or category structure suffered due to the difficulty of identifying appropriate words to distinguish the values of the differing feature dimensions. An additional potential problem with the stimuli in Experiment 7 was that three of the four non-hidden feature dimensions varied based on size such that the values were all effectively large/small. These descriptors potentially increased the cognitive capacity needed to distinguish between the feature dimensions and their values because of descriptor confusability and cross-dimension interference. The participants in Experiment 7 may not have had much cognitive capacity to spare due to this potential confusability of the feature value descriptors and the requirement for a full internal representation. The next experiment changed the stimuli to colour, shape and size dimensions so that each feature dimension and its values could be more easily and uniquely differentiated and described, thus

facilitating the ease of learning and potentially minimizing the use of suboptimal dimensional rules.

4.3. Experiment 8

4.3.1. Introduction

To reduce attempts to use rules to classify the typical and atypical instances in the category structure, Table 2 in Chapter 3, learning in phase 2 was changed to include continued presentation of the ordinary instances along with the typical and atypical instances. This inclusion of the ordinary instances meant that there was no longer any perfectly diagnostic rule that could be applied to the second learning phase.

To encourage participants to learn the attachment of the hidden features, a third learning phase had hidden feature inference trials for the typical and atypical instances mixed in with all of the previous classification trials. Additionally, this experiment changed the stimulus dimensions to make them more verbalizable and less confusable and so, easier to learn. Finally, this experiment included qualitative questions at the end to determine learning strategy by asking participants to describe what they used to guide their responding in the task.

4.3.2. Materials and Methods

4.3.2.1. Participants

128 Cardiff University Psychology students participated for course credit.

4.3.2.2. Materials and Procedure

Two stimuli feature dimensions and values were changed, see Figure 44, to improve learning: the size of the body band was changed to the colour of the body band, blue or green (and the size was set as the previous ‘short’ value). The size of the booster was changed to the length of the boosters with feature values of tall and short. Additionally, a second booster was included and both boosters were the previous ‘large’ size value. The wing size dimension with

wide/narrow feature values, the cone shape dimension with pointed/rounded feature values and the hidden features remained the same as in the previous experiment.

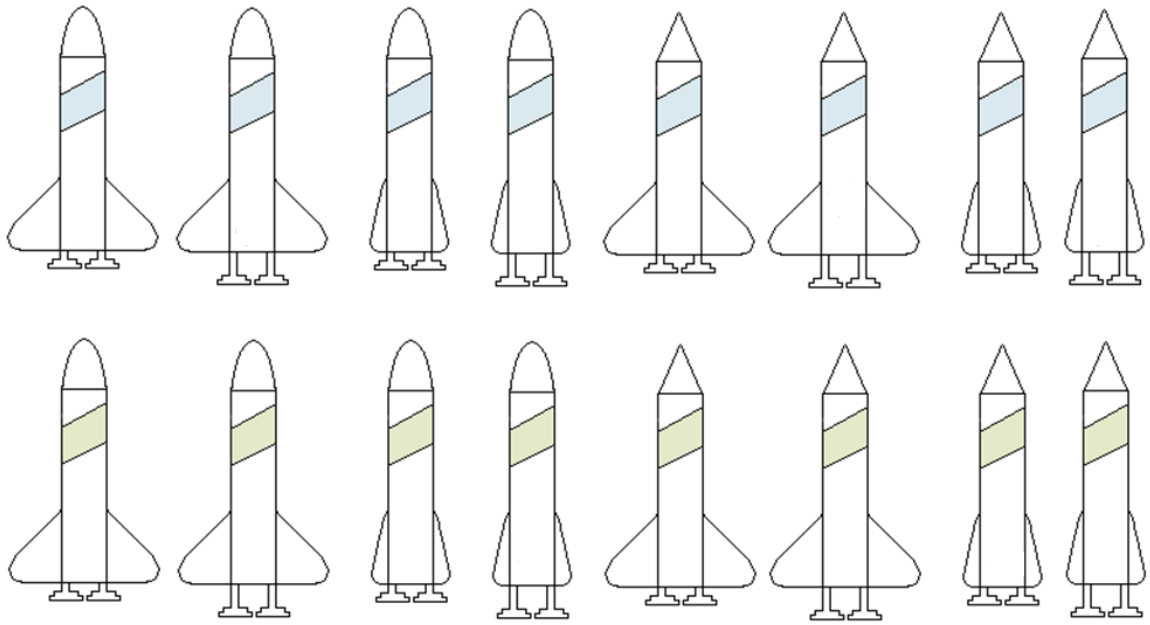

Figure 44. The 16 basic rocket ship stimuli used in Experiment 8, composed of binary features on four dimensions: nose cone shape (pointed/rounded), body band colour (blue/green), wing size (wide/narrow) and booster length (tall/short).

As in Experiment 7, participants learned the ordinary category instances first, in blocks 1-5. However, different from Experiment 7, the next blocks, blocks 6-10, had the typical and atypical instances (without their hidden features) mixed in with the ordinary instances. Finally, learning blocks 11-15 maintained the inclusion of the ordinary instances and showed the hidden features attached to the typical and atypical instances on the classification trials as well as mixing in feature inference trials querying those hidden features.

The testing phase eliminated the continuous generalization, the perceptual premise diversity and the perceptual inclusion fallacy trials from Experiment 7 as the new stimulus dimensions weren't all compatible with the continuous generalization trials. Additionally, the confidence rating trials for individual testing trials were removed. All other materials and procedures were the same as in Experiment 7 (though see Appendix B for minor changes to

one block of generalized classification trials). The full abstract structure of all training and testing blocks and average accuracies for each testing trial are in Appendix B.

4.3.3. Results

The average learning accuracy across all participants and blocks, the orange line in Figure 45, was only at 66%. So, a post data collection criterion was imposed of greater than 75% correct over the last learning block (block 15), the blue line in Figure 45, and this resulted in an average accuracy of 93% correct in that block. This criterion included about a third of the participants (43 out of 128) and most of the following analyses are based only on those participants who learned the task well unless otherwise stated.

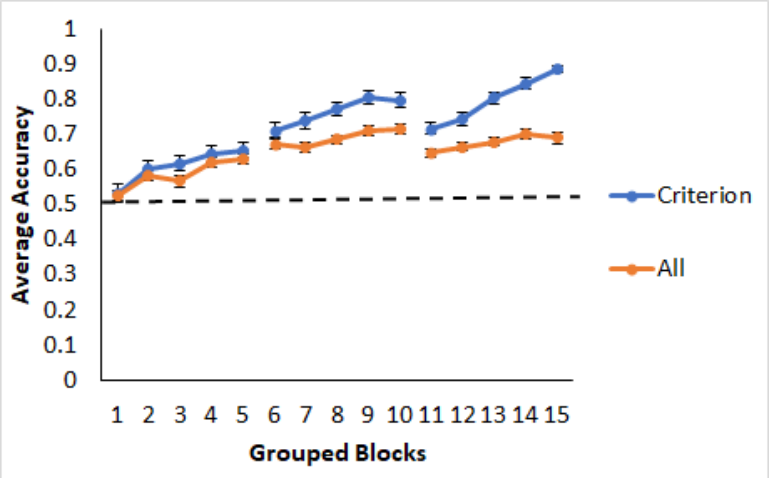

Figure 45. Average accuracy as proportion correct for all participants, orange line, and for learning criterion participants (greater than 75% accuracy in block 15), blue line, by learning block in Experiment 8. Blocks 1-5 were ordinary instance classification learning trials only, blocks 6-10 had typical and atypical instances without hidden features as well as ordinary instance classification learning trials, and blocks 11-15 had all classification typicality trials (with typical and atypical instance hidden features) as well as hidden feature inference learning trials. The three learning phase learning curves are separated by breaks in the lines. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

The learning phase data, Figure 46, show a typicality effect as significantly greater accuracy on the typical instances than the atypical in blocks 6-15 ($t(42) = 6.0$, $p < 0.001$). The ordinary trial accuracy across blocks 6-15 was also significantly less than the typical trial accuracy ($t(42) = 21.1$ $p < 0.001$) but not better than accuracy on the atypical instances ($t(42) = 2.4$, $p = 0.019$) in part because accuracy on the atypical instances was almost as high as the typical instances by the last block of training. Criterion participants achieved 89% percent correct overall in the last block on typical and atypical hidden feature inference, the blue and purple lines in Figure 46, suggesting good attachment of the hidden features to these instances. However, there was no significant difference in accuracy between inferring the typical and atypical hidden features in the learning phase ($t(42) = 0.4$, $p = 0.692$). In the two classification testing blocks, CL1 and CL2 on the far right in Figure 46, accuracy on the typical instances

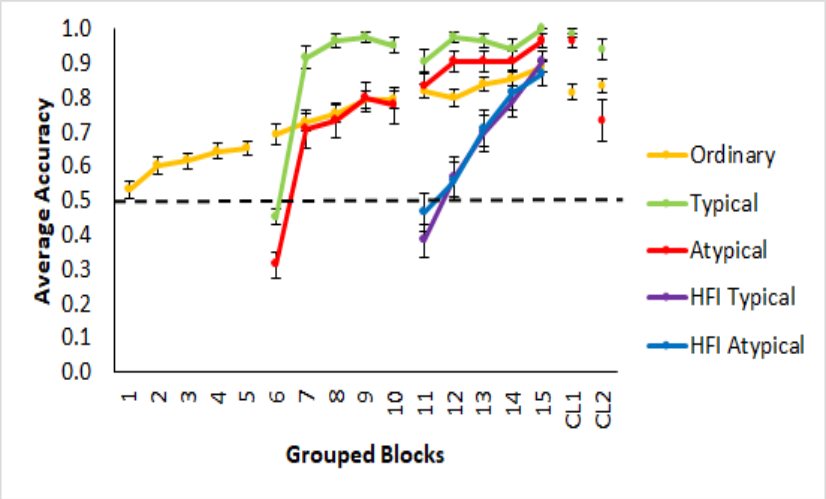

Figure 46. Averaged accuracy as proportion correct for all blocks of classification learning in Experiment 8, grouped by trial type--typical = green, ordinary = yellow, atypical = red, typical hidden feature inference = purple, atypical hidden feature inference = blue. Average accuracy in the two classification testing blocks, CL1 and CL2 is shown by dots on the far right for the different levels of typicality. The three learning phase learning curves are separated by breaks in the lines. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

was significantly higher than the atypical instances, ($t(42) = 3.4$, $p = 0.002$). Overall, there was a typicality effect, indicating sensitivity to the typicality structure with good attachment of the hidden features to that structure, but this didn't extend to a difference between typical and atypical hidden feature inference learning.

Improving on Experiment 7, accuracy on the hidden feature inferences, Figure 47 middle bar, was high ($t(42) = 13.3$, $p < 0.001$), indicating good attachment of the hidden features to the typical and atypical category instances. Further, the average accuracy over both blocks of the classification testing trials, Figure 47 left bar, was good, ($t(42) = 23.2$, $p < 0.001$) and there was a trend towards a preference for the atypical non-hidden features in the exception feature inference trials, Figure 47 right bar; however, this was not significant ($t(42) = 1.5$, $p = 0.129$). Overall, this suggests that participants maintained their learning into the testing phase including accurately attaching the hidden features to the typical and atypical instances, a necessary prerequisite for premise typicality.

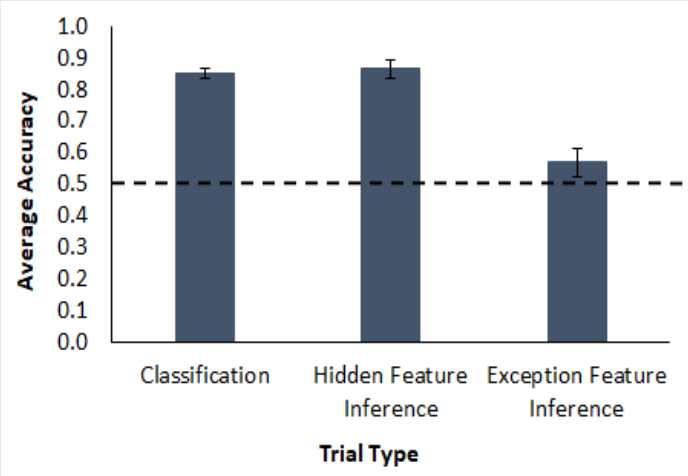

Figure 47. Average accuracy as proportion correct for the classification, hidden feature inference and exception feature inference testing trials in Experiment 8, grouped by trial type, see Table 3 in Chapter 3. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

However, despite strong attachment of the hidden features by the learning criterion participants, there was no effect of premise typicality, Figure 48 blue bars, for the generalized premise typicality trials ($t(42) = 0.1$, $p = 0.960$) or the ordinary premise typicality trials ($t(42) = 0.6$, $p = 0.553$). This suggests that participants had no preference for responding with the typical hidden feature over the atypical hidden feature for those trials.

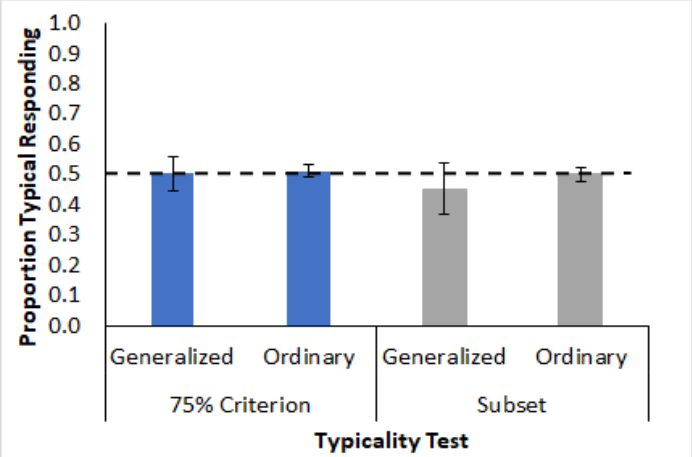

Figure 48. Average proportion of typical hidden feature responding by type of premise typicality trial, see Table 3 in Chapter 3, in Experiment 8. The blue bars are for the 75% learning criterion participants (greater than 75% correct in block 15) and the grey bars are for the subset criterion participants with greater than 75% correct in block 15 and the additional conditions of 85% correct over all hidden feature inference testing trials and an overall typicality effect. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

As this experiment had a large number of participants it is possible to select only those participants who strongly met all of the identified prerequisites for premise typicality by showing greater than 75% accuracy in the final block of learning *and* greater than 85% accuracy in attaching hidden features to the instances in testing and, once these participants were selected, further analyses showed an overall typicality effect: accuracy was significantly higher for the typical instances compared to the ordinary instances ($t(28) = 3.2$, $p = 0.003$) and the ordinary instances had marginally higher accuracy than the atypical instances ($t(28) = 1.8$,

p = 0.076). These participants are arguably the most likely to show premise typicality in terms of both learning well but also in terms of having typicality-based differences in performance. There were 29 participants who met these stringent criteria; however they showed no effect of premise typicality, Figure 48 grey bars, for the generalized premise typicality trials (t(28) = 0.6, p = 0.575) or the ordinary premise typicality trials (t(28) = 0, p = 1, note the average response proportion was literally 0.5).

The classic paradigm version of premise typicality, Figure 49, resulted in significantly higher likelihood ratings for the argument that included the typical item as a premise compared to the argument with the atypical premise for all participants (t(127) = 4.4, p < 0.001) and for those who met the 75% learning criterion, (t(42) = 3.6, p = 0.001). So, these participants did show a classic premise typicality effect despite the lack of premise typicality in the perceptual learning task.

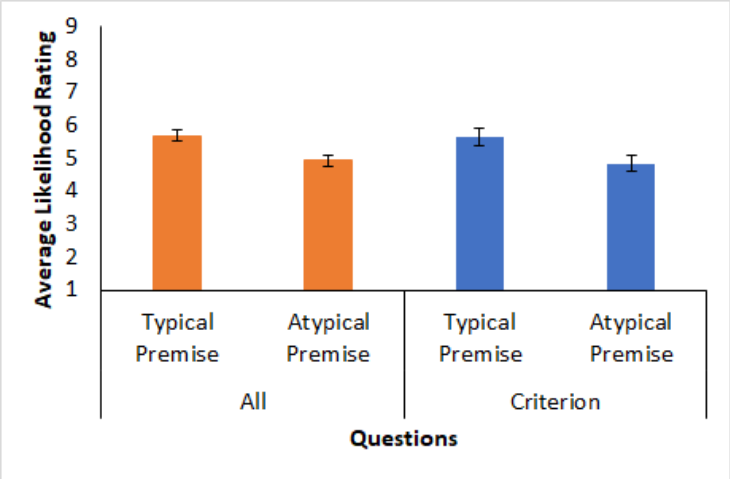

Figure 49. Average argument likelihood ratings for the classic paradigm premise typicality testing trials in Experiment 8. Orange bars = all data, blue bars = 75% criterion data. Error bars show ±1 standard error.

The error diagrams help to clarify the strategy used by each participant, see Figure 50 (and for a more detailed explanation see p. 81 and pp. 105-107). As in the previous experiment, each row corresponds to a specific category instance and every column corresponds to a

learning block. As not all learning trials occurred in all learning blocks with later blocks phasing in additional trials, the rows towards the bottom of each individual participant's error diagram indicate performance on trials introduced in later blocks and are arranged such that all trials in a given block are grouped together in one column. So, blocks (columns) 6-10 added the typical and atypical classification trials and blocks (columns) 11-15 added the hidden feature inference learning trials for the typical and atypical instances. The "Good HF" grouping from the last experiment was eliminated here as there was no clear group of participants who learned the typical and atypical instances well but the ordinary instances poorly. As shown in Figure 50, 23% of participants were apparently using a dimensional rule, somewhat less than the 38% of the prior experiment, but the prevalence of dimensional rules still suggests the task was quite hard. Further, the large group of poor learners (45%) also indicate that this task was hard to learn even with the updated stimuli. However, there were a nontrivial number of good learners, (32%).

At the end of this experiment, participants were asked to write down what rule or other learning method they used during the tasks. The experimenter coded these written descriptions and assigned each participant to a strategy type, see Figure 51. The 'Optimal rule' group was specified in terms of descriptions of an accurate unidimensional rule plus individual exemplar exceptions. This group consisted of participants specifying between two and four exceptions which could correspond to optimal performance in the first learning phase or in both the first and second learning phases. The 'Suboptimal rule' group was specified as any unidimensional, configural or any idiosyncratic rule that corresponded to inaccuracies in performance. The 'Exemplars' group was specified as a description of more than four individual instances (so as to distinguish from optimal rule users specifying exceptions) or also specified as a mention of

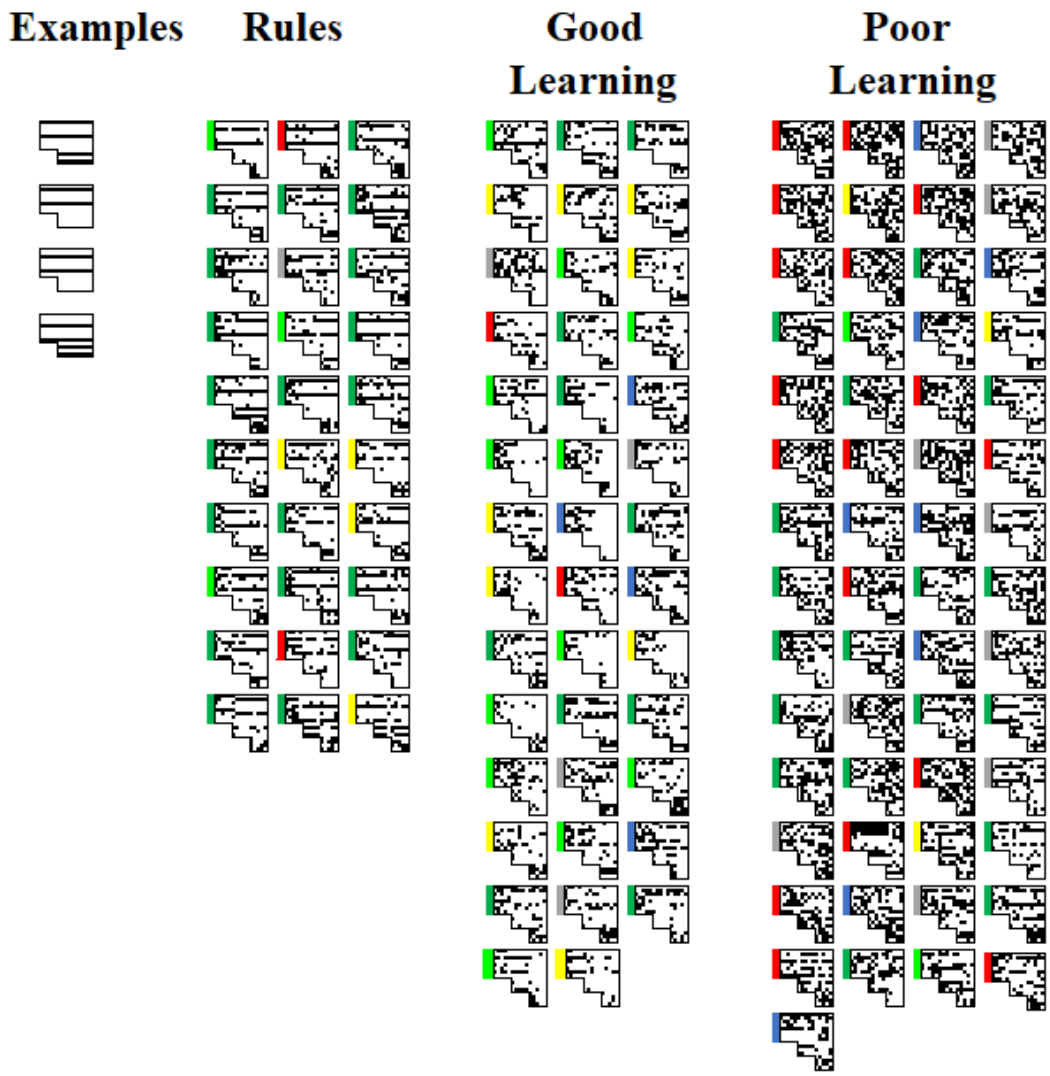

Figure 50. Error diagrams showing performance of each individual participant on each trial over classification and feature inference learning in Experiment 8. Instances are arranged in rows and testing blocks are arranged in columns. 'Examples' of the predictions of all unidimensional rules are shown on the left side in order with a dimension one rule at the top and a dimension four rule at the bottom for comparison. The 'Examples' graphs exclude the hidden feature inference trials as unidimensional rules do not make predictions for these trials. Coloured blocks to the left of each diagram indicate the experimenter identified learning strategy based on participants' qualitative self-reports: light green = optimal rule, dark green = suboptimal rule, yellow = prototypes, blue = exemplars, grey = ambiguous, red = no rule/poor reported learning.

an instance memorization strategy. The 'Prototypes' group was specified as a description of three or more features indicated to 'mostly' belong to one category. The 'Ambiguous' group was specified in terms of responses that lacked sufficient information and/or were unclear. Finally, participants were coded into the 'No rule/poor reported learning' group if they indicated that they were not able to learn the task. Participant reports were consistent with the stimuli they saw for 61% of participants, ambiguous for 35% of participants, and inconsistent for only 4% of participants, see Figure 51. This suggests that participants self-reports were fairly but not extremely accurate. Overall, the self-report results indicate that about a third of participants used dimensional rules, roughly consistent with the number of apparent dimensional rule users in the error diagrams.

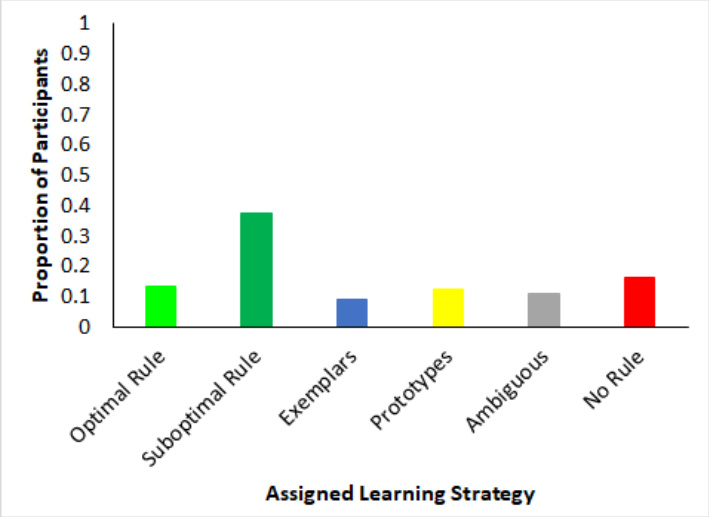

Figure 51. Proportion of participants in each of the 6 strategy groups: light green = optimal rule, dark green = suboptimal rule, blue = exemplars, yellow = prototypes, grey = ambiguous, red = no rule/poor reported learning.

Just as there was no effect of premise typicality, there was no premise typicality *like* effect on the blank feature inference trials, Figure 52 right bar, i.e., no preference for responding with the typical hidden feature over the atypical, ($t(42) = 0.7$, $p = 0.519$). However, the premise conclusion similarity testing trials, Figure 52 left bar, resulted in a strong preference for

responding with the typical hidden feature, ($t(42) = 12.7$, $p < 0.001$). This indicates a premise typicality *like* effect that is nonetheless based on similarity rather than typicality.

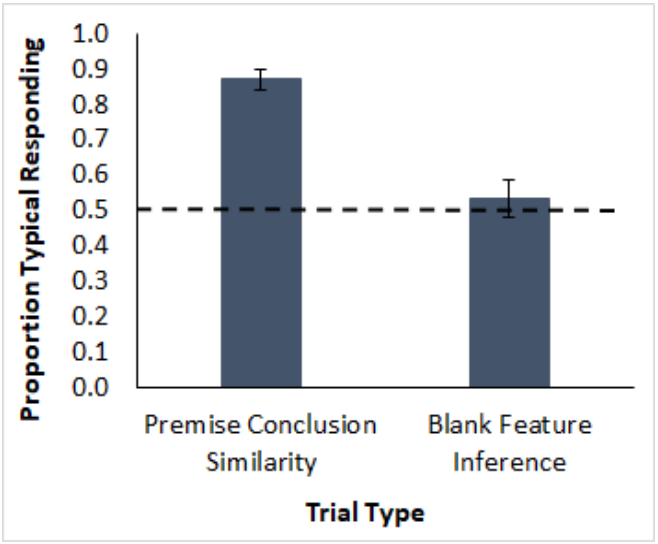

Figure 52. Average proportion typical hidden feature responding for premise conclusion similarity and blank feature inference testing trials grouped by trial type, see Table 3 in Chapter 3, in Experiment 8. The dashed line is a reference for two-option chance responding. Error bars show ±1 standard error.

4.3.4. Discussion

In this experiment accuracy was high by the end of the learning phase when a criterion was applied. Additionally, a strong effect of typicality occurred with higher accuracy on the typical instances than the atypical instances and participants were able to accurately attach the hidden features to each of the typical and atypical instances. However, there was no effect of premise typicality. The error diagrams suggest that roughly half of the participants who learned something about the category structure were responding consistent with a dimensional rule. Further, a subset of very good learners who met all of the prerequisites of learning and testing still showed no effect of premise typicality despite a learning criterion higher than the accuracy predicted by the use of suboptimal unidimensional rules. This suggests that the lack of a premise typicality effect in the perceptual paradigm was not simply due to a predominance of

unidimensional rule use in learning, however, a tendency to attend to only a subset of feature dimensions cannot be completely eliminated as a potential explanation for the observed lack of premise typicality.

Participants reported their own strategies in self-report data reasonably well as only 4% of participants reported a strategy inconsistent with their performance. Rule use was the dominant strategy that participants reported using, and the error diagrams also suggest that a little under half of all participants who learned were using rules. Whilst the subset data using the stringent prerequisite based criterion suggests caution in attributing the lack of premise typicality to suboptimal dimensional rule use, the fact that half of the participants who learned were likely using dimensional rules and such rules do not reflect a clear appreciation of the typicality structure leaves the reason for the absence of the effect uncertain: it may be due to a lack of typicality appreciation or a lack of sensitivity and power for detecting a small premise typicality effect.

4.4. General Discussion

In Experiment 7 the learning task introduced the ordinary instances in early learning blocks and the typical and atypical instances in a second learning phase. This resulted in good end of learning performance on the typical and atypical instances, however, that didn't translate into accurate feature inference of the hidden features for the typical and atypical instances, and there was no effect of premise typicality. The good performance on the typical and atypical instances was potentially due to the presence of a perfectly diagnostic unidimensional rule in the second phase of learning that allowed 100% accuracy on the typical and atypical instances. This rule could have been based on non-hidden feature dimensions two or three and did not require participants to learn about the relations of each instance with the hidden features. This may have led to poor testing performance on the hidden feature inference trials and the subsequent lack of premise typicality. Ultimately, the prominent use of unidimensional rules

suggests that participants found the learning task difficult and the rates of learning could be improved.

To improve learning and reduce dimensional rule use, Experiment 8 changed the stimuli and intermingled ordinary instances with the typical and atypical instances in the second learning phase to reduce dimensional rule use in that phase based on the lack of a perfectly valid rule. Additionally, Experiment 8 included feature inferences based on the hidden features to improve the learning of those features. This had the intended effect with high accuracy in attaching the hidden features and the strengthening/maintaining of the typicality and classification prerequisites. However, the participants who learned the task in terms of the 75% learning criterion (and therefore supposedly had internally represented the category) showed no effect of premise typicality and neither did a subset of participants who learned the task well and met the prerequisites strongly.

Overall the results of these two experiments are consistent with three possible conclusions. The first possible conclusion is that the prerequisites have not been satisfied with sufficient power to detect the effect. Second, the stated prerequisites may be necessary for premise typicality, but they are not sufficient and there is some further requirement for this effect in the perceptual category learning paradigm. Third, premise typicality doesn't actually exist in the perceptual paradigm after controlling for the similarity of the test cases to the typical and atypical category instances.

The first possible conclusion is that premise typicality didn't occur due to the difficulty of learning the instances of this variant of the family resemblance structure, resulting in few participants meeting the learning criterion and lessening the power to find a premise typicality effect. An obvious solution is to train participants for far longer in the present tasks, perhaps across multiple sessions. This would likely improve performance, but it wouldn't necessarily eliminate rule-based learning strategies or produce an appreciation of typicality differences.

Another possible solution is to use other category structures that participants may find easier to learn such as structures with reduced feature dimensions (Zeithamova & Maddox, 2006). I have done several (unreported) experiments based on reduced dimensions attempting to do this which resulted in good learning by a majority of participants but did not show a premise typicality effect. Also, an information integration category structure (Ashby & Gott, 1988) might be sufficient to disrupt dimensional rule use and produce better appreciation of typicality.

The second possible conclusion is that the lack of premise typicality is due to a methodological failure to satisfy some additional requirement for a premise typicality effect. One possible new prerequisite might be that the hidden feature needs to be attached to a subcategory rather than a specific instance as a direct analogue to the classic categorical induction paradigm. In the classic paradigm the key components, e.g. robins and penguins, are both subcategories of the category bird rather than specific instances of the category bird, unlike in the current experiment. Whilst this is methodologically possible in the perceptual paradigm, it would be practically difficult. In particular it would require a more complex learning task with presumably substantially more training than the current experiments to obtain good learning.

The final possible reason for the failure to observe premise typicality is that it doesn't actually exist after controlling for similarity in the perceptual paradigm and perhaps even in the classical categorical induction paradigm. Experiments 4-6 and 8 all resulted in a preference for responding with the typical hidden feature when the test item was more similar to the typical instance than to the atypical but not when the test item was equally similar to the typical and atypical instances i.e. there was an effect of premise conclusion similarity but not premise typicality. However, fully controlling for similarity in the classic paradigm with real-world categories is at best, extremely difficult.

While all three conclusions are plausible given the current set of results, the first conclusion needs to be more compellingly eliminated, i.e. the prerequisites need to be met with sufficient power via a larger number of participants. The ideal follow up experiment would do this whilst also clarifying the possible additional prerequisite stated above that the key comparison needs to be made on subcategories rather than specific instances. Such an experiment would include a hierarchy of categories learned in great detail, potentially over multiple learning sessions, by a large number of participants trained to a very high learning criterion. If this experiment still did not demonstrate a premise typicality effect, the third conclusion relating to the impact of similarity on the premise typicality effect would become more relevant.

It is important to note the expectation from these data had these experiments been conducted without reference to the classic categorical induction paradigm. Typicality effects have been widely demonstrated in the literature and these effects alone should have elicited a preference for responding consistent with the features associated with a typical instance (Light et al., 1979; Lin et al., 1990; McCloskey & Glucksberg, 1978; Medin & Schaffer, 1978; Nosofsky, 1988; Rosch & Mervis, 1975; Rosch et al., 1976; Rothbart & Lewis, 1988; Spalding & Murphy, 1999; etc.). Although the hidden features being induced were less visually frequent than the non-hidden features, they were still associated with the typical instance and short of finding a premise typicality effect, there should still have been a preference for the typical features to the extent that typicality effects are pervasive. From this perspective, premise typicality effects really should exist in perceptual categorization.

Overall these findings demonstrate an initial attempt to find premise typicality via feature inference testing following classification learning of a family resemblance structure. Though learning of the structures was good by at least a subset of participants, no effect of premise typicality occurred. This result in combination with the failure to find premise

typicality in the category summary based decision-making task (see Chapter 3) was surprising. The difficulty of finding categorical induction effects in the perceptual categorization domain begins to cast doubt on the existence of such effects, but additional experimental designs with higher power will be needed to definitively establish whether premise typicality effects are actually distinct from premise conclusion similarity effects.

Chapter Five - General Discussion

The ability to make inferences about instances of a category is a crucial cognitive ability that allows efficient interaction with everyday reality. There is little consensus on the nature of the category representation underlying feature inference (see Chapter 1, p. 16) and the overarching purpose of this thesis was to assess feature inference in perceptual categorization, motivated by its adaptive importance for categories. This research has evaluated attribute induction in terms of feature inference learning—learning about category instances by making feature inferences with feedback—and feature inference decision-making—making feature inferences for known category instances.

The aim of Experiments 1-3 was to assess the category representations resulting from feature inference learning of the classic category structures from Shepard et al. (1961). They constructed six 'types' of category structure that differed in the complexity of the rule characterizing them and in their learning difficulty with Type I as the easiest, Type II was more difficult, Types III, IV and V were harder still but were as equally difficult as each other and Type VI was the most difficult i.e. I<II<III=IV=V<VI. Type I was learnable by a unidimensional rule on the first feature dimension and Type II was learnable by a configural rule across the first two feature dimensions. As Types III and IV are not perfectly learnable by feature inference on the first feature dimension (see Chapter 2, p. 29) and Types III, IV and V are equally difficult, Experiments 1 and 2 only assessed Type V which was learnable by a unidimensional rule plus the memorization of two exception instances. Finally, Type VI was learnable by complete memorization of instances or by the Odd-Even rule in which one instance was memorized and if a further instance differed from that instance on one or three features it belonged to a different category and if it varied by two features it belonged to the same category as the memorized instance. Participants learned these types by classification or feature inference.

Experiments 1-2 tested a label-bias hypothesis that participants in the feature inference conditions would preferentially try to form rules based on the category labels in contrast to classification learning, in which rules could not be based on the category label (due to the absence of the label as part of the stimuli). The tendency to form label-based rules in feature inference manifested in a Type I feature inference learning advantage over classification. In the Type I feature inference condition, the unidimensional rule that allowed good learning of the category structure included the category label and many feature inference learners achieved perfect accuracy almost immediately in the learning phase consistent with forming a label-based rule. In classification learning by contrast, the correct feature dimension to use to form a unidimensional rule was initially no more likely to be chosen than rules based on the other two feature dimensions and so across participants, finding the correct unidimensional rule took longer, resulting in an early accuracy advantage in the feature inference condition. The results of Type V feature inference in Experiment 1 also supported the label bias hypothesis in terms of a tendency to perseverate on a label-based rule despite persistent errors on the two exceptions to that rule; a pattern that didn't occur in Type V classification learning. Additionally, the lack of differentiation in the feature inference learning curves for Types II, V and VI is consistent with similar attempts to use suboptimal label-based rules across the more complex types.

The error diagrams for Experiments 1-3 showed performance on each individual trial for each individual participant and indicated relatively sudden improvements in accuracy consistent with the acquisition of a rule. Qualitative data of participants' self-reported learning strategies in Experiments 2-3 indicated that a relatively high proportion of participants were using rules in that they specified a correct rule for the type they were learning and the stimuli they saw. Overall, the results of Experiments 1-3 indicated the dominance of rule representation in both classification and feature inference learning but greater consistency in rule strategy in feature inference consistent with the label-bias hypothesis.

A further interesting finding from Experiment 1 was that participants found Types II, V and VI quite a bit harder to learn than, for example, the participants in the standard replication in Nosofsky et al. (1994), apparently due to differences in the stimuli. Only 13% of participants in the Type VI feature inference learning condition learned the structure by the end of training. Improving the verbalizability of the stimuli and reducing their confusability in Experiment 2 increased the proportion of participants who learned but learning was still worse than Nosofsky et al. (1994), see Figures 4 and 9 in Chapter 2. Experiment 3 further assessed the impact of verbalizability by comparing the rocket ship stimuli from Experiment 2 to the classic stimuli from Shepard et al. (1961). The rocket ship stimuli were line drawings with binary features that varied on dimensions including the size of the wings, the colour of a band drawn across the body of the ship and the shape of a cone positioned at the top of the ship (see Figure 7 in Chapter 1). The classic stimuli were one of two shapes, circle or triangle, that varied in size and colour. Experiment 3 found that the classic stimuli were substantially easier to learn than the rocket ship stimuli and corresponded to more compact verbal rules. This suggests that the classic Shepard et al. (1961) stimuli are extremely specialized in terms of allowing inordinately compact verbal rules, even compared to the generally simple stimuli commonly used in perceptual categorization tasks.

Experiments 4-8 evaluated categorical induction as a reasoning process used to infer features/attributes for members of categories. The standard categorical induction paradigm (Chapter 1, pp. 20-24) asks participants to rate argument strength in terms of generalizing the presence of a hidden feature from one category member to another, for example from sparrows to geese. There are many influences on these argument likelihood ratings, but the most important one for the purposes of this thesis is premise typicality; the more typical the premise category instance, the stronger the judged argument that is based on it compared to an argument based on the atypical category instance, e.g. “Sparrows have property X Therefore Geese have

property X" is judged to be a stronger argument than, "Penguins have property X Therefore Geese have property X." (Hayes et al., 2010). The purpose of Experiments 4-8 was to produce an analogue of premise typicality in the perceptual categorization paradigm using feature inference decision-making so as to be able to assess the category representation underlying feature inference using standard exemplar and prototype models (Homa, 1984; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Posner & Keele, 1968; Reed, 1972; Smith & Minda, 1998 etc.). The feature inference task was based on categories with a strong typicality structure, i.e. with a typical and an atypical instance, and 'hidden' features attached to the typical and atypical instances, (see Chapter 3, Figure 18). Participants were tested on a new instance that was equally similar to the typical and atypical instances and were asked which hidden feature the new instance had: the one attached to the typical instance or the one attached to the atypical instance. This assessed an analogue of the classic premise typicality effect in the perceptual categorization paradigm across Experiments 4-8.

As classic tests of premise typicality are based on summary descriptions, e.g. 'Robins have a sesamoid bone', Experiments 4-6 were based on a summary presentation of perceptual instances in two categories with hidden features attached to the typical and atypical instances. Experiment 4 used a basic summary decision-making task to assess premise typicality using feature inference testing. Participants were able to correctly categorize the instances and were significantly more accurate in classifying the typical category instances compared to the atypical instances, thus showing a typicality effect. However, participants did not show a premise typicality effect. This was arguably because they were not able to correctly attach the hidden features to the typical and atypical instances. A plausible reason for the lack of hidden feature attachment was a lack of engagement with the category summary due to a potential tendency for guessing induced by testing trials without a clear correct answer immediately before the key tests of premise typicality.

To enhance engagement and reduce guessing, Experiment 5 included more testing trials with clear correct answers in terms of the category summary. This maintained classification accuracy, and as before, participants showed a typicality effect, but importantly they also showed reasonably good attachment in terms of accurately inferring hidden features. However, there was still no effect of premise typicality, despite the presence of a typicality effect. But this typicality effect may have been due to pseudo-typicality brought about by averaging across participants who were using unidimensional rules, as seen in the error diagrams. Specifically, there were four unidimensional rules that a participant could have used, none of which resulted in errors on the typical instances. Each of these rules produced an error on one of the ordinary category instances and thus reduced ordinary instance accuracy somewhat compared to typical instance accuracy. Finally, there were two rules which produced errors on the atypical instances, and thus reduced atypical instance accuracy even further compared to typical instance accuracy. Overall, the use of dimensional rules across participants would result in lower accuracy on the atypical instance because more rules are inconsistent with its correct classification than for the typical instance. So, although the data showed a typicality effect across participants, individual participants may not have appreciated the typicality structure of the categories.

Experiment 6 reduced dimensional rule use by having participants complete a learning task based on the category summary that asked them to attend to every feature and dimension of the ordinary category instances before the key test of premise typicality. Participants in Experiment 6 accurately classified the category instances and showed an effect of typicality despite the almost complete elimination of dimensional rule users. In addition, they accurately attached the hidden features to the typical and atypical instances; nevertheless, there was still no effect of premise typicality.

Experiments 4-6 identified three prerequisites for a premise typicality effect such that the failure to establish any one prerequisite provided a plausible explanation for the lack of an effect. The first was that participants needed to be sufficiently engaged with the category summary so as to accurately classify each instance into the correct category. The second was that participants needed to show an appreciation of the typicality structure within each category, for example, in terms of higher accuracy on the typical instance than the atypical instance. The third prerequisite was that participants needed to accurately attach the hidden features to their respective typical and atypical category instances for the key trials to be a test of premise typicality. These three prerequisites were all reasonably well satisfied in Experiment 6 and yet there was not even a tendency towards a premise typicality effect. However, it is possible that the premise typicality effect failed to occur because of the summary-based methodology; Experiments 4-6 used a summary presentation of the category instances, so the category was externally represented rather than internally, mentally represented, and in the classic categorical induction paradigm, the categories are usually internally represented. Internal category representations might plausibly be more likely to produce a premise typicality effect due to forgetting and interference between features that make the typical features more prominent in the representation. While summary based decision-making tasks are common and ecologically plausible, the categories in categorical induction tasks are usually real-world categories with rich internal representations, e.g. robins and birds. This suggests the possibility of a fourth prerequisite for premise typicality that participants must have an internalized, learned representation of the categories.

Experiment 7 removed the category summary and replaced it with a trial-by-trial classification learning task with feedback for all the category instances. Learning occurred in two phases to strengthen the typicality effect. The phased design first presented the ordinary category instances. Following this, the second phase presented only the typical and atypical

category instances with their hidden features. Participants correctly classified the instances (prerequisite 1) and showed a typicality effect in terms of greater accuracy in initially classifying the typical instances over the atypical instances (prerequisite 2), however, participants did not learn the attachment of the hidden features to the typical and atypical instances (prerequisite 3). So as in experiment 4, this lack of attachment provides a plausible reason for the observed lack of premise typicality. The error diagrams showed a group of participants who correctly classified the typical and atypical instances but did not correctly classify the ordinary instances. The performance of this group is consistent with the application of a perfectly diagnostic unidimensional rule during the second phase of learning which could, in particular, be used without attending to or learning about the hidden features.

This perfectly diagnostic rule was only applicable to the typical and atypical instances when they were learned separately from the ordinary instances. So, to preclude the use of this kind of rule, Experiment 8 maintained training of the ordinary instances into the second phase along with the typical and atypical instances (without hidden features present) so that participants learned the typical and atypical instances in relation to the ordinary instances rather than via a separate rule. In addition, a third learning phase continued training on the ordinary, typical and atypical instances (with hidden features present) while also training hidden feature attachment by feature inference. Participants who met a learning criterion and therefore formed a strong internal representation of the categories (prerequisite 4) correctly classified the category instances (prerequisite 1), showed a typicality effect (prerequisite 2) and accurately attached the hidden features to the typical and atypical instances (prerequisite 3); nevertheless there was still no effect of premise typicality. Further, a subset of participants who met the prerequisites especially strongly also did not show premise typicality.

Overall, Experiments 4-8 didn't find any evidence of premise typicality. This might mean that the stated prerequisites were not met with sufficient power to detect an effect;

however the novelty of the present tests of premise typicality and their methodological difference from the classic paradigm make it difficult to conceptually specify power in a formal sense as it is not clear what the effect size in this paradigm should be. Nor is it straightforward to specify a minimum effect size such that smaller values, even if corresponding to significant effect, would not constitute a conceptually meaningful indication of premise typicality. As such even Bayesian support for a null hypothesis isn't completely straightforward in this context. Nevertheless, Experiment 8 had at least a moderate level of intuitive power by the standards of perceptual category learning. Alternatively, the observed lack of premise typicality might mean that there is an additional prerequisite that needs to be met to find a premise typicality effect. A possibility is that a structure with subcategories within categories needs to be used in the perceptual paradigm to make a closer comparison with the classic categorical induction paradigm. Finally, the lack of premise typicality might mean that premise typicality does not exist in the perceptual categorization paradigm when similarity is controlled for. That is, similarity may be fundamentally responsible for replications of classic premise typicality despite attempts to control for it. Regardless of which explanation is correct, and it may be a combination of them, having failed to establish an analogue of premise typicality using perceptual categories, this research has not facilitated a direct comparison of representation models. However, qualitative data in Experiment 8 and error diagrams for Experiments 1-3, 5 and 7-8 suggest that rules were a prominent strategy in feature inference learning and decision-making, and many participants tried to use a rule-based representation.

Experiments 1-2 show the dominance of the category label in feature inference of perceptual categories as argued by Yamauchi and Yu (2008), specifically the category label when it represents category membership information. Similarly, both Gelman and Markman (1986) and Yamauchi and Markman (2000) found that the category label was used to guide feature inference for category instances above perceptual similarity in both children and adults.

Yamauchi and Markman (1998) hypothesized that feature inference learning promotes prototype representation due to a focus on learning within category information compared to classification learning which focuses on between category information. And the current results are somewhat consistent with the spirit of this hypothesis, that classification and feature inference learning produce differences in the category representation. In the Shepard et al. (1961) category structures, both feature inference and classification learning encouraged the formation of rules and thus rule-based representation. However, in feature inference learning this representation was centered on the category labels and corresponded to a tendency for a within category focus, a subtle difference from the representation for classification learning which was not centered on the labels.

In feature inference testing of a family-resemblance structure seemingly well suited to typicality-based effects and therefore strongly suited to prototype representation, there was no clear evidence of premise typicality effects despite the evidence for typicality effects. So those results do not support prototype representation though they aren't necessarily evidence against it either. However, note that the qualitative data suggested that rule use was a reasonably predominant representation strategy. Therefore, these experiments are consistent with the conclusions of Johansen and Kruschke (2005) that feature inference learning encourages the learning of label-based rules that can mimic a prototype model but is not necessarily prototype representation. Overall, the prevalence of rule use in Experiments 1-3, 5 and 7-8, as seen from the error diagrams and qualitative data indicate at least a tendency to try to use rules for representing perceptual categories.

It is worth clarifying the relationship between the present evidence for the attempted formation of rules and the rejected Classical View (Chapter 1, pp. 1-3) that categories have necessary and sufficient features for classifying an instance as a member of a category. The generally agreed fact that most real-world categories do not have necessary and sufficient

conditions does not fit especially comfortably with the evidence for attempted rule use, but this is also not completely inconsistent as the attempted rules need not be optimal in the way that the Classical View would suggest. And category labels as a basis for feature inference can be similar to prototypes, so the rejection of the Classical View of categories need not be inconsistent with a bias to form label-based, though potentially suboptimal rules. Additionally, the continued success of the COVIS framework (Ashby et al., 1998) with a module based on rules suggests the continued relevance of rule representation.

A key limitation of Experiments 7-8 is that the variant of the family resemblance structure used in the learning task proved hard for participants to learn in terms of satisfying all of the identified prerequisites. While Experiment 8 had a reasonably large sample size of 128 participants, once the prerequisites were strongly applied, 29 participants remained. While this is a reasonable sample size, a larger sample size might help clarify the absence of premise typicality. Another possible improvement would be to train participants in the current task for longer. However, the error diagrams show that many participants learned essentially nothing by the end of training, so more training might not actually produce all that much more learning. Another alternative might be to specify a category typicality structure based on a smaller number of feature dimensions that are easier to learn. However, this might also have the effect of actually encouraging rules at the expense of typicality perception.

The largest potential limitation of Experiments 4-8 is the possibility that the test of perceptual premise typicality is not equivalent to the specification of the test in the classic paradigm. Specifically, the classic paradigm uses categories within categories e.g. the category bird includes the subcategory ‘robin’ but robin itself is a category based on lots of instances of robins. In contrast, Experiments 4-8 used single instances, a typical instance and an atypical instance, rather than typical and atypical subcategories of instances as a basis for tests of premise typicality. It might be possible to create an experimental design with perceptual

subcategories in perceptual categories. However, participants already struggled to learn the current, less complex category structure so this may not be all that practical. And more importantly, even if the current design is not an exact match to the classic version of premise typicality, it is doing the same thing in spirit and so should still produce an analogue of the effect. The fact that there was not even a tendency towards responding with the typical feature, leaves the question of why not? One possible reason why there was no tendency towards responding with the typical feature is that these experiments so carefully controlled for test item similarity to the typical and atypical category instances, something which is far more difficult to do for real-world categories due to their complexity. Potentially, similarity effects could be influencing typicality effects in real-world categories.

The results for Experiments 4-8 are consistent with the idea that similarity rather than typicality is the basis for premise typicality even in the classic categorical induction paradigm. The premise typicality and premise conclusion similarity tests are equivalent in as much as they require responding with the typical hidden feature over the atypical hidden feature. However, the comparison between the premises and conclusion is equated in similarity for the premise typicality tests but not for the premise conclusion similarity tests. When the test item was more similar to the typical instance in the premise conclusion similarity tests, typical hidden feature responding was significantly higher than chance, an effect that did not occur for premise typicality when the test item was equally similar to the typical and atypical instances. This suggests that responding consistent with a premise typicality effect occurs based on perceptual categories when typicality is confounded with similarity. This seems to support exemplars more than prototypes as, if they were forming a representation based on prototypes, they should have shown premise typicality. Future experiments may be able to control for the effects of similarity in the classic versions of categorical induction tasks more carefully to tease apart typicality and similarity. In the classic paradigm, the comparisons between the premises

and conclusion control for similarity through participants ratings of similarity, however these ratings may not reflect the internal representation of the similarity between a premise and conclusion, so distinguishing typicality and similarity for anything other than fairly controlled perceptual categories may prove challenging.

More speculatively, there are subtle ways in which the current methodologies can be argued to be biased against prototypes but also biased in their favour. The majority of participants in Experiments 7-8 did not learn the categories well, despite the fact that they were linearly separable and hence prototype representable. So, the difficulty of learning doesn't particularly support prototype representation as a default, or the categories should have been relatively easy to learn, at least all else being equal. Also, a tendency for prototype representation should presumably have made the typical hidden features quite a bit easier to learn than the atypical hidden features; the typical hidden feature was attached to the prototypical instance, but prototype representation by itself doesn't have anything to attach the atypical hidden feature to. That might have manifested in a premise typicality effect as a result of substantially greater accuracy in attaching the typical hidden feature to the typical instance than the atypical hidden feature to the atypical instance (because participants should not be able to learn the atypical instance very well). However, the results of Experiment 8 didn't show a clear difference in accuracy for the hidden features and nor did they show premise typicality. But it is also possible to argue the opposite; the fact that participants were asked to learn both the typical and atypical hidden features effectively drove them away from their default, prototype representation. And the difficulty of learning argument can also be reversed: it is possible that the difficulty in learning the current category structures with instances composed of only a small number of feature dimensions may reflect the unrepresentativeness of such category structures relative to real-world categories composed of instances with a much larger number of feature dimensions. So rather than simplifying the categories, this suggests that

further experiments might try substantially increasing the number of semi-diagnostic feature dimensions. Ultimately these results haven't provided the clear evidence for prototypes that they might have, but the remaining possible explanations for these results also haven't definitely ruled out prototype representation.

In summary, this thesis attempted to assess the category representation underlying feature inference learning of the classic Shepard et al. (1961) category structures and feature inference decision-making in a perceptual categorical induction paradigm. This was based on the idea (summarized by Murphy, 2002) that prototype representation is a plausible basis for categorical induction effects and, by extension, feature inference learning and decision-making. However, none of the experiments here provided compelling evidence for prototype representation, but rather provided tentative support for category label-based rule representation underlying feature inference learning and decision-making.

References

Ahn, W. K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41*(4), 361-416.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-429.

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30*(1), 119-128.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442-481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33-53.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*(3), 372-400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154-179.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 638-648.

Chin-Parker, S. (2011). What Varying the Learning Task and Category Structure Reveals About Inference Learning. *In Proceedings of the Annual Meeting of the Cognitive Science Society, 33*(33), 625-630.

Conant, M. B., & Trabasso, T. (1964). Conjunctive and disjunctive concept formation under equal-information conditions. *Journal of Experimental Psychology, 67*(3), 250-255.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *The Quarterly Journal of Experimental Psychology, 65*(3), 439-464.

Dopkins, S., & Gleason, T. (1997). Comparing exemplar and prototype models of categorization. *Canadian Journal of Experimental Psychology*, *51*(3), 212-230.

Edmunds, C., & Wills, A. J. (2016). Modeling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*(38), 69-74.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*(2), 107-168.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183-209.

Gratton, C., Evans, K. M., & Federmeier, K. D. (2009). See what I mean? An ERP study of the effect of background knowledge on novel object processing. *Memory & Cognition*, *37*(3), 277-291.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68-107.

Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 576-595.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 441-461.

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition, 12*(2), 151-164.

Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley interdisciplinary reviews: Cognitive science, 1*(2), 278-292.

Haygood, R. C., & Bourne, L. E., Jr. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review, 72*(3), 175-195.

Heit, E. (1998). Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 712-731.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*(4), 569-592.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In *Psychology of Learning and Motivation* (Vol. 39, pp. 163-199). Academic Press.

Honke, G., Conaway, N., & Kurtz, K. (2016). Switch it up: Learning categories via feature switching. *In Proceedings of the Annual Meeting of the Cognitive Science Society, 38*(38), 2693–2698.

Homa, D. (1984). On the nature of categories. In *Psychology of Learning and Motivation* (Vol. 18, pp. 49-94). Academic Press.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7*(6), 418-439.

Howell (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thompson/Wadsworth.

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological Monographs, 28*(1), i-86.

Hunt, E. B., & Hovland, C. I. (1960). Order of consideration of different types of concepts. *Journal of Experimental Psychology, 59*(4), 220-225.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child: Classification and Seriation.* London: Routledge & Kegan Paul.

Jee, B. D., & Wiley, J. (2014). Learning about the internal structure of categories through classification and feature inference. *The Quarterly Journal of Experimental Psychology, 67*(9), 1786-1807.

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General, 126*(3), 248-277.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1433-1458.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning?. *Cognitive Psychology, 45*(4), 482-553.

Johansen, M. K., Savage, J., Fouquet, N., & Shanks, D. R. (2015). Salience not status: How category labels influence feature inference. *Cognitive Science*, *39*(7), 1594-1621.

Kalish, C. W., & Gelman, S. A. (1992). On wooden pillows: Multiple classification and children's category-based inductions. *Child Development*, *63*(6), 1536-1557.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 829-846.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22-44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review, 14*(4), 560-576.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and

Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 552-572.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 754-770.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review*, *3*(1), 95-99.

Lewandowsky, S. (2011). Working memory capacity and categorization: individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 720-738.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1666-1684.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87-137.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 212-228.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance, 23*(4), 1153-1169.

Lin, P. J., Schwanenflugel, P. J., & Wisenbaker, J. M. (1990). Category typicality, cultural familiarity, and the development of category knowledge. *Developmental Psychology, 26*(5), 805-813.

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43*(2), 283-297.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*(4), 829-835.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review, 111*(2), 309-332.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077-1083.

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 539-555.

Markman, A. B., & Maddox, W. T. (2003). Classification of exemplars with single-and multiple-feature manifestations: The effects of relevant dimension variation and category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 107-117.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets?. *Memory & Cognition, 6*(4), 462-472.

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*(1), 128-148.

McNorgan, C., Kotack, R. A., Meehan, D. C., & McRae, K. (2007). Feature-feature causal relations and statistical co-occurrences in object concepts. *Memory & Cognition, 35*(3), 418-431.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(1), 37-50.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 333-352.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome?. *Cognitive Psychology*, *32*(1), 49-96.

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition*, *99*(3), 237-273.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355-368.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241-253.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242-279.

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: a comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1518-1533.

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*(3), 775-799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 275-292.

Mooney, R. J. (1993). Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning*, *10*(1), 79-110.

Murphy, G. (2004). *The big book of concepts.* MIT press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 904-919.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*(2), 148-193.

Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When do people integrate across categories?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 263-276.

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54-65.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352-369.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of " multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, *7*(3), 375-402.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 211-233.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, *3*(2), 222-226.

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

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*(4), 247-255.

Nosofsky, R. M., & 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*(5), 924-940.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185-200.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review, 6*(3), 495-503.

Palmeri, T. J., & Blalock, C. (2000). The role of background knowledge in speeded perceptual categorization. *Cognition, 77*(2), B45-B57.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology Section A, 54*(1), 197-235.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 416-432.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353-363.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*(1), 1-14.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 811-828.

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

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 347-363.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*(3), 264-314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*(3), 323-360.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition, 91*(2), 113-153.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1-41.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 659-683.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10*(4), 759-784.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 665-681.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou and A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 21-59). Cambridge: Cambridge University Press.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin*, *127*(6), 827-852.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573-605.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495-553.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491-502.

Rothbart, M., & Lewis, S. (1988). Inferring category attributes from exemplar attributes: Geometric shapes and social categories. *Journal of Personality and Social Psychology*, *55*(6), 861-872.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*(5), 919-932.

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, *18*(3), 229-239.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 641-649.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325-345.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1-42.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121*(3), 278-304.

Smith, J. D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science, 13*(5), 437-442.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts (Vol. 9)*. Cambridge, MA: Harvard University Press.

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*(6), 1411-1436.

Smith, J. D., & Minda, J. P. (2001). Journey to the center of the category: the dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 984-1002.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398-414.

Smith, E. E., Osherson, D. N., Rips, L. J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, *12*(4), 485-527.

Smith, J. D., Redford, J. S., & Haas, S. M. (2008). Prototype abstraction by monkeys (Macaca mulatta). *Journal of Experimental Psychology: General*, *137*(2), 390-401.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, *49*(1-2), 67-96.

Smoke, K. L. (1932). An objective study of concept formation. *Psychological Monographs, 42*(4), i-46.

Spalding, T. L., & Murphy, G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(2), 525-538.

Spalding, T. L., & Murphy, G. L. (1999). What is learned in knowledge-related categories? Evidence from typicality and feature frequency judgments. *Memory & Cognition*, *27*(5), 856-867.

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1251-1263.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, *42*(1), 51-73.

Sweller, N., & Hayes, B. K. (2010). More than one kind of inference: Re-examining what's learned in feature inference and classification. *The Quarterly Journal of Experimental Psychology*, *63*(8), 1568-1589.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Verbeemen, T., Vanpaemel, W., Pattyn, S., Storms, G., & Verguts, T. (2007). Beyond exemplars and prototypes as memory representations of natural concepts: A clustering approach. *Journal of Memory and Language*, *56*(4), 537-554.

Vitkin, A. Z., Coley, J. D., & Hu, R. (2005). Children's use of relevance in open-ended induction in the domain of biology. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *27*(27), 2319-2324.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, *15*(3), 630-637.

Ward, T. B., Vela, E., & Hass, S. D. (1990). Children and adults learn family-resemblance categories analytically. *Child Development, 61*(3), 593-605.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*(2), 158-194.

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 449-468.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*(2), 221-281.

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

Yamauchi, T., Kohn, N., & Yu, N. Y. (2007). Tracking mouse movement in feature inference: Category labels are different from feature labels. *Memory & Cognition*, *35*(5), 852-863.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 585-593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124-148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776-795.

Yamauchi, T., & Yu, N. Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition, 36*(3), 544-553.

Yu, N. Y., Yamauchi, T., & Schumacher, J. (2008). Rediscovering symbols: The role of category labels in similarity judgment. *Journal of Cognitive Science*, *9*(2), 89-109.

Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: a reassessment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1160-1173.

Žauhar, V., Bajšanski, I., & Domijan, D. (2014). Metacognitive Monitoring of Rule-Based Category Learning Tasks. *Proceedings of the Trieste Symposium on Perception and Cognition,* 162-165.

Žauhar, V., Bajšanski, I., & Domijan, D. (2016). Concurrent dynamics of category learning and metacognitive judgments. *Frontiers in Psychology*, *7*, 1-11.

Zeigler, D. E., & Vigo, R. (2018). Classification errors and response times over multiple distributed sessions as a function of category structure. *Memory & Cognition*, *46*(7), 1041-1057.

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

Ziori, E., & Dienes, Z. (2008). How does prior knowledge affect implicit and explicit concept learning?. *The Quarterly Journal of Experimental Psychology*, *61*(4), 601-624.

Appendices

7.1. Appendix A: All Experiment 2 testing trials on the second and third stimulus dimensions.

Testing Trials

Type I	*Type II*	*Type V*	*Type VI*
A1?1	A111	A1?1	A111
A1?1	A110	A110	A010
A1?0	A001	A1?1	A001
A1?0	A000	A000	A100
B0?1	B011	B0?1	B011
B0?1	B010	B0?1	B110
B0?0	B101	B010	B101
B0?0	B100	B100	B000
A11?	A11?	A11?	A111
A10?	A11?	A11?	A010
A11?	A00?	A101	A001
A10?	A00?	A000	A100
B01?	B01?	B01?	B011
B00?	B01?	B001	B110
B01?	B10?	B01?	B101
B00?	B10?	B100	B000

7.2. Appendix B: Full specification of all trials in Experiments 4-8 with average response proportions for each trial.

For the first structure table for each experiment, the abstract category structure used is in the left, top corner. The next column is the descriptor for the construct that each block was training/testing followed in the next column by the abstract structure of the trials. The next two columns contain the average response proportions over all participants with the first column showing averaged abstract correct/typical/label-based responding depending on the trial. If the trial had a unique correct answer in the category summary or given in the learning, then the first column was a measure of responding with that correct answer. If the trial was querying either a hidden feature for an instance other than the typical and atypical of each category or a non-hidden feature and there was no correct answer (as in there was no exact match or multiple matches in the category summary or learning task) column one was a measure of responding with the typical feature. If the trial was comparing the effects of the label against another stimulus component, the feature typical of the category the label is denoting is considered to be label-based responding therefore, for these trials, column one represents the proportion of participants responding with the label consistent option. The second column shows the averaged abstract incorrect/atypical/hidden feature-based responding which is the opposite of the responding displayed in the first column. It shows responding with the incorrect answer, the atypical feature or the feature typical of the category denoted by the hidden feature respectively. The final two columns report the response proportions based on participants who met a criterion of greater than 75% correct in the classification testing block (Experiments 4-6) or greater than 75% in either blocks 17-20 (Experiment 7) or block 15 (Experiment 8). For each trial in the table, the letter/number/symbol in red was queried. When a letter or number was marked in red, this indicated that when this label or feature was queried there was a unique, correct answer in the category summary or learning task. A red question mark indicated that

when that label or feature was queried there was no basis in the category summary or learning task for responding with any one answer over the other or there were two answers that were consistent with the information provided on that trial.

Experiment 4

Abstract Structure	Trial Type	Testing Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A 3111_ _	*Classification*	A3111_ _	0.838	0.162	1.000	0.000
A 1311_ _		A1311_ _	0.676	0.324	1.000	0.000
A 1131_ _		A1131_ _	0.838	0.162	0.778	0.222
A 1113_ _		A1113_ _	0.892	0.108	1.000	0.000
A 1111V_		A1111_ _	0.973	0.027	1.000	0.000
A 3113 X_		A3113_ _	0.730	0.270	1.000	0.000
B 1333_ _		B1333_ _	0.865	0.135	1.000	0.000
B 3133_ _		B3133_ _	0.649	0.351	1.000	0.000
B 3313_ _		B3313_ _	0.757	0.243	0.889	0.111
B 3331_ _		B3331_ _	0.865	0.135	1.000	0.000
B 3333_Y		B3333_ _	0.973	0.027	1.000	0.000
B 1331_Z		B1331_ _	0.676	0.324	1.000	0.000
	Generalized Classification	?1133_ _	0.568	0.432	0.444	0.556
		?3311_ _	0.351	0.649	0.333	0.667
		?1313_ _	0.432	0.568	0.556	0.444
		?3131_ _	0.622	0.378	0.556	0.444
	Generalized Premise Typicality	A1133?_	0.432	0.568	0.444	0.556
		B3311_?	0.351	0.649	0.556	0.444
		A1313?_	0.649	0.351	0.889	0.111
		B3131_?	0.595	0.405	0.444	0.556
	Original Premise Typicality	A3111?_	0.297	0.703	0.333	0.667
		A1113?_	0.784	0.216	0.889	0.111
		B1333_?	0.243	0.757	0.222	0.778
		B3331_?	0.784	0.216	0.889	0.111
	Premise Conclusion Similarity	A1311?_	0.405	0.595	0.333	0.667
		A1131?_	0.892	0.108	1.000	0.000
		B3133_?	0.432	0.568	0.333	0.667
		B3313_?	0.892	0.108	1.000	0.000
	Hidden Feature Inference	A1111V_	0.432	0.568	0.444	0.556
		A3113X_	0.595	0.405	0.556	0.444
		B3333_Y	0.486	0.514	0.333	0.667
		B1331_Z	0.568	0.432	0.444	0.556
	Ambiguous	A?111_ _	0.486	0.514	0.222	0.778
		B?333_ _	0.405	0.595	0.333	0.667
		A111?_ _	0.703	0.297	0.556	0.444
		B333?_ _	0.649	0.351	0.556	0.444
	Exception Feature Inference	A3113X_	0.757	0.243	0.667	0.333
		A3113X_	0.757	0.243	0.889	0.111
		B1331_Y	0.703	0.297	0.556	0.444
		B1331_Y	0.838	0.162	1.000	0.000

Continued

Label vs Feature	A3?33_Y	0.378	0.622	0.556	0.444
	B1?11_V	0.378	0.622	0.667	0.333
Continuous Generalization	A2112?_	0.757	0.243	0.556	0.444
	B1221_?	0.297	0.703	0.222	0.778
	A0110?_	0.703	0.297	0.778	0.222
	B4114_?	0.649	0.351	0.556	0.444
	A1221?_	0.865	0.135	0.667	0.333
	B3223_?	0.919	0.081	0.889	0.111
	A1001?_	0.324	0.676	0.444	0.556
	B3443_?	0.243	0.757	0.222	0.778
Blank Feature Inference	A_ _ _ _?_	0.703	0.297	0.667	0.333
	B_ _ _ _ _?	0.811	0.189	1.000	0.000
Label vs Hidden Features	A_?_ _ _Y	0.459	0.541	0.556	0.444
	A_ _?_ _Z	0.541	0.459	0.556	0.444
	B_?_ _V_	0.541	0.459	0.667	0.333
	B_ _?_X_	0.568	0.432	0.778	0.222
Premise Diversity	A2212?_	0.270	0.730	0.222	0.778
	B2232_?	0.514	0.486	0.444	0.556
The Inclusion Fallacy	A_ _ _ _?_	0.514	0.486	0.444	0.556
	A3003?_	0.162	0.838	0.111	0.889
	B_ _ _ _ _?	0.757	0.243	0.889	0.111
	B1441_?	0.351	0.649	0.111	0.889

Experiment 5

Abstract Structure	Trial Type	Testing Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A 3111_ _	*Classification*	A3111_ _	0.875	0.125	0.944	0.056
A 1311_ _		A1311_ _	0.854	0.146	0.889	0.111
A 1131_ _		A1131_ _	0.708	0.292	1.000	0.000
A 1113_ _		A1113_ _	0.854	0.146	0.944	0.056
A 1111V_		A1111_ _	1.000	0.000	1.000	0.000
A 3113 X_		A3113_ _	0.729	0.271	0.944	0.056
B 1333_ _		B1333_ _	0.896	0.104	1.000	0.000
B 3133_ _		B3133_ _	0.792	0.208	0.889	0.111
B 3313_ _		B3313_ _	0.646	0.354	0.889	0.111
B 3331_ _		B3331_ _	0.854	0.146	1.000	0.000
B 3333_Y		B3333_ _	0.917	0.083	1.000	0.000
B 1331_Z		B1331_ _	0.688	0.313	1.000	0.000
	Classification With Hidden Features Block 1	A3111_ _	0.875	0.125	1.000	0.000
		A1311_ _	0.875	0.125	0.833	0.167
		A1131_ _	0.854	0.146	1.000	0.000
		A1113_ _	0.854	0.146	0.889	0.111
		A1111V_	0.979	0.021	1.000	0.000
		A3113X_	0.625	0.375	0.778	0.222
		B1333_ _	0.854	0.146	1.000	0.000
		B3133_ _	0.854	0.146	0.833	0.167
		B3313_ _	0.708	0.292	1.000	0.000
		B3331_ _	0.875	0.125	0.944	0.056
		B3333_Y	0.958	0.042	1.000	0.000
		B1331_Z	0.729	0.271	0.944	0.056
	Ordinary Premise Typicality	A3111?_	0.458	0.542	0.389	0.611
		A1113?_	0.354	0.646	0.278	0.722
		B1333_?	0.521	0.479	0.500	0.500
		B3331_?	0.458	0.542	0.389	0.611
	Hidden Feature Inference Block 1	A1111?_	0.958	0.042	1.000	0.000
		A3113?_	0.917	0.083	1.000	0.000
		B3333_?	0.917	0.083	0.889	0.111
		B1331_?	0.833	0.167	0.944	0.056
	Generalized Premise Typicality	A1133?_	0.500	0.500	0.500	0.500
		B3311_?	0.542	0.458	0.500	0.500
		A1313?_	0.396	0.604	0.167	0.833
		B3131_?	0.771	0.229	0.778	0.222
	Hidden Feature Inference Block 2	A1111?_	0.958	0.042	1.000	0.000
		A3113?_	0.958	0.042	1.000	0.000
		B3333_?	0.917	0.083	0.944	0.056
		B1331_?	0.896	0.104	0.944	0.056
	Classification With Hidden Features Block 2	A3111_ _	0.875	0.125	0.944	0.056
		A1311_ _	0.917	0.083	0.944	0.056
		A1131_ _	0.854	0.146	0.889	0.111
		A1113_ _	0.896	0.104	0.944	0.056
		A1111V_	0.958	0.042	1.000	0.000
		A3113X_	0.667	0.333	0.889	0.111
		B1333_ _	0.938	0.063	1.000	0.000
		B3133_ _	0.896	0.104	0.944	0.056
		B3313_ _	0.813	0.188	0.944	0.056
		B3331_ _	0.854	0.146	0.889	0.111
		B3333_Y	0.938	0.063	0.944	0.056
		B1331_Z	0.750	0.250	0.944	0.056
	Premise Conclusion Similarity	A1311?_	0.938	0.063	1.000	0.000
		A1131?_	0.917	0.083	1.000	0.000
		B3133_?	0.958	0.042	1.000	0.000
		B3313_?	0.958	0.042	1.000	0.000
	Hidden Feature Inference Block 3	A1111?_	0.938	0.063	1.000	0.000
		A3113?_	0.917	0.083	0.889	0.111
		B3333_?	0.958	0.042	1.000	0.000
		B1331_?	0.896	0.104	0.944	0.056
	Hidden Feature Inference Block 4	A1111V_	0.938	0.063	1.000	0.000
		A3113X_	0.917	0.083	0.944	0.056
		B3333_Y	0.958	0.042	1.000	0.000
		B1331_Z	0.917	0.083	0.944	0.056

Continued

Hidden Feature Inference Block 5	A1111?_	0.979	0.021	1.000	0.000
	A3113?_	0.938	0.063	0.944	0.056
	B3333_?	0.875	0.125	0.889	0.111
	B1331_?	0.896	0.104	0.944	0.056
Ambiguous	A?111_ _	0.729	0.271	0.556	0.444
	B?333_ _	0.729	0.271	0.611	0.389
	A111?_ _	0.729	0.271	0.667	0.333
	B333?_ _	0.708	0.292	0.556	0.444
Exception Feature Inference	A3113X_	0.938	0.063	1.000	0.000
	A3113X_	0.917	0.083	0.889	0.111
	B1331_Y	0.917	0.083	1.000	0.000
	B1331_Y	0.833	0.167	0.889	0.111
Classification With Hidden Features Block 3	A3111_ _	0.813	0.188	1.000	0.000
	A1311_ _	0.854	0.146	0.944	0.056
	A1131_ _	0.854	0.146	0.944	0.056
	A1113_ _	0.833	0.167	0.944	0.056
	A1111V_	0.958	0.042	1.000	0.000
	A3113X_	0.625	0.375	0.778	0.222
	B1333_ _	0.875	0.125	1.000	0.000
	B3133_ _	0.938	0.063	0.944	0.056
	B3313_ _	0.750	0.250	0.833	0.167
	B3331_ _	0.813	0.188	0.889	0.111
	B3333_Y	0.938	0.063	1.000	0.000
	B1331_Z	0.604	0.396	0.722	0.278
Label vs Feature	A3?33_Y	0.458	0.542	0.389	0.611
	B1?11_V	0.604	0.396	0.556	0.444
Continuous Generalization	A2112?_	0.500	0.500	0.389	0.611
	B1221_?	0.438	0.563	0.556	0.444
	A0110?_	0.479	0.521	0.611	0.389
	B4114_?	0.500	0.500	0.500	0.500
	A1221?_	0.583	0.417	0.611	0.389
	B3223_?	0.563	0.438	0.556	0.444
	A1001?_	0.563	0.438	0.500	0.500
	B3443_?	0.563	0.438	0.611	0.389
Blank Feature Inference	A_ _ _ _?_	0.521	0.479	0.444	0.556
	B_ _ _ _ _?	0.604	0.396	0.500	0.500
Label vs Hidden Features	A_?_ _ _Y	0.479	0.521	0.444	0.556
	A_ _?_ _Z	0.604	0.396	0.611	0.389
	B_?_ _V_	0.438	0.563	0.444	0.556
	B_ _?_X_	0.625	0.375	0.667	0.333
Generalized Classification	?1133_ _	0.417	0.583	0.444	0.556
	?3311_ _	0.417	0.583	0.500	0.500
	?1313_ _	0.542	0.458	0.444	0.556
	?3131_ _	0.479	0.521	0.389	0.611
Generalized Classification with Hidden Features	?1133_ _	0.542	0.458	0.556	0.444
	?3311_ _	0.521	0.479	0.556	0.444
	?1313_ _	0.458	0.542	0.333	0.667
	?3131_ _	0.542	0.458	0.389	0.611
Premise Diversity	A2212?_	0.438	0.563	0.500	0.500
	B2232_?	0.375	0.625	0.222	0.778
The Inclusion Fallacy	A_ _ _ _?_	0.417	0.583	0.333	0.667
	A3003?_	0.438	0.563	0.444	0.556
	B_ _ _ _ _?	0.771	0.229	0.889	0.111
	B1441_?	0.375	0.625	0.333	0.667
Typical Premise		4.625	4.778		
Atypical Premise		3.458	3.833		
Typical Conclusion		4.500	4.444		
Atypical Conclusion		3.958	4.111		
More Diverse Premise		4.917	4.722		
Less Diverse Premise		4.938	4.167		
Category Conclusion		5.313	4.944		
Instance Conclusion		3.354	2.889		
Category Premise		7.146	6.500		
Instance Premise		5.688	7.222		

Experiment 6 – Summary Learning Task

Abstract Structure	Trial Type	Testing Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A 3111_ _	*Classification Learning*	A3111	0.750	0.250	0.758	0.242
A 1311_ _		A1311	0.896	0.104	0.939	0.061
A 1131_ _		A1131	0.792	0.208	0.879	0.121
A 1113_ _		A1113	0.792	0.208	0.848	0.152
A 1111V_		B1333	0.813	0.188	0.939	0.061
A 3113X_		B3133	0.729	0.271	0.818	0.182
B 1333_ _		B3313	0.792	0.208	0.879	0.121
B 3133_ _		B3331	0.813	0.188	0.848	0.152
B 3313_ _	*Feature Inference Dimension 1 Learning*	A3111	0.521	0.479	0.636	0.364
B 3331_ _		A1311	0.833	0.167	0.818	0.182
B 3333_Y		A1131	0.729	0.271	0.848	0.152
B 1331_Z		A1113	0.729	0.271	0.697	0.303
		B1333	0.521	0.479	0.636	0.364
		B3133	0.771	0.229	0.788	0.212
		B3313	0.667	0.333	0.788	0.212
		B3331	0.667	0.333	0.758	0.242
	Feature Inference Dimension 2 Learning	A3111	0.875	0.125	0.848	0.152
		A1311	0.625	0.375	0.788	0.212
		A1131	0.750	0.250	0.788	0.212
		A1113	0.667	0.333	0.758	0.242
		B1333	0.771	0.229	0.848	0.152
		B3133	0.521	0.479	0.576	0.424
		B3313	0.792	0.208	0.909	0.091
		B3331	0.708	0.292	0.727	0.273
	Feature Inference Dimension 3 Learning	A3111	0.854	0.146	0.939	0.061
		A1311	0.813	0.188	0.818	0.182
		A1131	0.542	0.458	0.636	0.364
		A1113	0.750	0.250	0.788	0.212
		B1333	0.792	0.208	0.848	0.152
		B3133	0.854	0.146	0.939	0.061
		B3313	0.521	0.479	0.667	0.333
		B3331	0.792	0.208	0.848	0.152
	Feature Inference Dimension 4 Learning	A3111	0.729	0.271	0.788	0.212
		A1311	0.750	0.250	0.788	0.212
		A1131	0.854	0.146	0.909	0.091
		A1113	0.479	0.521	0.606	0.394
		B1333	0.792	0.208	0.818	0.182
		B3133	0.854	0.146	0.909	0.091
		B3313	0.750	0.250	0.848	0.152
		B3331	0.458	0.542	0.576	0.424

Experiment 6 – Testing Trials

Abstract Structure	Trial Type	Testing Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A 3111_ _	***Classification***	A3111_ _	0.958	0.042	0.970	0.030
A 1311_ _		A1311_ _	0.875	0.125	0.970	0.030
A 1131_ _		A1131_ _	0.833	0.167	0.909	0.091
A 1113_ _		A1113_ _	0.938	0.063	1.000	0.000
A 1111V_		A1111_ _	0.896	0.104	1.000	0.000
A 3113X_		A3113_ _	0.854	0.146	1.000	0.000
B 1333_ _		B1333_ _	0.917	0.083	1.000	0.000
B 3133_ _		B3133_ _	0.792	0.208	0.970	0.030
B 3313_ _		B3313_ _	0.854	0.146	0.970	0.030
B 3331_ _		B3331_ _	0.958	0.042	1.000	0.000
B 3333_Y		B3333_ _	0.896	0.104	0.970	0.030
B 1331_Z		B1331_ _	0.813	0.188	0.939	0.061
	Classification with Hidden Features Block 1	A3111_ _	0.896	0.104	0.970	0.030
		A1311_ _	0.792	0.208	0.879	0.121
		A1131_ _	0.792	0.208	0.909	0.091
		A1113_ _	0.833	0.167	0.909	0.091
		A1111V_	0.896	0.104	0.909	0.091
		A3113X_	0.708	0.292	0.818	0.182
		B1333_ _	0.917	0.083	0.939	0.061
		B3133_ _	0.854	0.146	0.970	0.030
		B3313_ _	0.875	0.125	0.970	0.030
		B3331_ _	0.833	0.167	0.970	0.030
		B3333Y_	0.917	0.083	0.970	0.030
		B1331Z_	0.667	0.333	0.727	0.273
	Ordinary Premise Typicality	A3111?_	0.542	0.458	0.485	0.515
		A1113?_	0.458	0.542	0.545	0.455
		B1333_?	0.438	0.563	0.455	0.545
		B3331_?	0.563	0.438	0.485	0.515
	Hidden Feature Inference Block 1	A1111?_	0.833	0.167	0.848	0.152
		A3113?_	0.875	0.125	0.879	0.121
		B3333_?	0.833	0.167	0.939	0.061
		B1331_?	0.792	0.208	0.848	0.152
	Generalized Premise Typicality	A1133?_	0.646	0.354	0.606	0.394
		B3311_?	0.542	0.458	0.545	0.455
		A1313?_	0.458	0.542	0.485	0.515
		B3131_?	0.438	0.563	0.455	0.545
	Hidden Feature Inference Block 2	A1111?_	0.896	0.104	0.939	0.061
		A3113?_	0.854	0.146	0.909	0.091
		B3333_?	0.917	0.083	0.909	0.091
		B1331_?	0.875	0.125	0.909	0.091
	Classification with Hidden Features Block 2	A3111_ _	0.875	0.125	0.909	0.091
		A1311_ _	0.771	0.229	0.909	0.091
		A1131_ _	0.875	0.125	0.939	0.061
		A1113_ _	0.854	0.146	0.970	0.030
		A1111V_	0.875	0.125	0.909	0.091
		A3113X_	0.688	0.313	0.758	0.242
		B1333_ _	0.813	0.188	0.909	0.091
		B3133_ _	0.771	0.229	0.879	0.121
		B3313_ _	0.813	0.188	0.909	0.091
		B3331_ _	0.875	0.125	0.909	0.091
		B3333Y_	0.813	0.188	0.818	0.182
		B1331Z_	0.708	0.292	0.818	0.182
	Premise Conclusion Similarity	A1311?_	0.833	0.167	0.818	0.182
		A1131?_	0.729	0.271	0.667	0.333
		B3133_?	0.667	0.333	0.697	0.303
		B3313_?	0.708	0.292	0.758	0.242
	Hidden Feature Inference Block 3	A1111?_	0.875	0.125	0.909	0.091
		A3113?_	0.938	0.063	0.970	0.030
		B3333_?	0.896	0.104	0.970	0.030
		B1331_?	0.813	0.188	0.909	0.091

Continued

Hidden Feature Inference Block 4	A1111V_	0.875	0.125	0.909	0.091
	A3113X_	0.938	0.063	0.970	0.030
	B3333_Y	0.875	0.125	0.909	0.091
	B1331_Z	0.896	0.104	0.970	0.030
Hidden Feature Inference Block 5	A1111?_	0.875	0.125	0.939	0.061
	A3113?	0.938	0.063	0.939	0.061
	B3333_?	0.896	0.104	0.879	0.121
	B1331_?	0.792	0.208	0.848	0.152
Ambiguous	A?111_ _	0.396	0.604	0.303	0.697
	B?333_ _	0.313	0.688	0.242	0.758
	A111?_ _	0.417	0.583	0.364	0.636
	B333?_ _	0.438	0.563	0.364	0.636
Exception Feature Inference	A3113X_	0.833	0.167	0.879	0.121
	A3113X_	0.750	0.250	0.848	0.152
	B1331_Y	0.813	0.188	0.848	0.152
	B1331_Y	0.833	0.167	0.879	0.121
Classification with Hidden Features Block 3	A3111_ _	0.875	0.125	0.970	0.030
	A1311_ _	0.771	0.229	0.879	0.121
	A1131_ _	0.875	0.125	0.939	0.061
	A1113_ _	0.792	0.208	0.879	0.121
	A1111V_	0.854	0.146	0.879	0.121
	A3113X_	0.646	0.354	0.788	0.212
	B1333_ _	0.854	0.146	0.909	0.091
	B3133_ _	0.875	0.125	0.970	0.030
	B3313_ _	0.792	0.208	0.848	0.152
	B3331_ _	0.771	0.229	0.909	0.091
	B3333Y_	0.854	0.146	0.879	0.121
	B1331Z_	0.646	0.354	0.788	0.212
Label vs Feature	A3?33_Y	0.500	0.500	0.455	0.545
	B1?11_V	0.563	0.438	0.515	0.485
Continuous Generalization	A2112?_	0.521	0.479	0.515	0.485
	B1221_?	0.542	0.458	0.545	0.455
	A0110?_	0.438	0.563	0.424	0.576
	B4114_?	0.479	0.521	0.455	0.545
	A1221?_	0.542	0.458	0.485	0.515
	B3223_?	0.500	0.500	0.455	0.545
	A1001?_	0.396	0.604	0.394	0.606
	B3443_?	0.417	0.583	0.333	0.667
Blank Feature Inference	A_ _ _ _?_	0.625	0.375	0.545	0.455
	B_ _ _ _ _?	0.479	0.521	0.545	0.455
Label vs Hidden Features	A_?_ _ _Y	0.583	0.417	0.545	0.455
	A_ _?_ _Z	0.542	0.458	0.576	0.424
	B_?_ _ W_	0.563	0.438	0.545	0.455
	B_ _?_X_	0.542	0.458	0.515	0.485
Generalized Classification	?1133	0.583	0.417	0.606	0.394
	?3311_ _	0.604	0.396	0.636	0.364
	?1313_ _	0.354	0.646	0.303	0.697
	?3131_ _	0.458	0.542	0.515	0.485
Generalized Classification with Hidden Features	?1133_ _	0.500	0.500	0.455	0.545
	?3311_ _	0.646	0.354	0.697	0.303
	?1313_ _	0.438	0.563	0.424	0.576
	?3131_ _	0.438	0.563	0.394	0.606
Premise Diversity	A2212?_	0.396	0.604	0.394	0.606
	B2232_?	0.521	0.479	0.515	0.485
The Inclusion Fallacy	A_ _ _ _?_	0.500	0.500	0.394	0.606
	A3003?_	0.396	0.604	0.364	0.636
	B_ _ _ _ _?	0.688	0.313	0.848	0.152
	B1441_?	0.542	0.458	0.396	0.604
Typical Premise		4.73	4.64		
Atypical Premise		4.06	3.94		
Typical Conclusion		4.88	5.30		
Atypical Conclusion		3.90	4.06		
More Diverse Premise		5.21	5.30		
Less Diverse Premise		5.23	4.67		
Category Conclusion		5.21	4.82		
Instance Conclusion		3.83	3.73		
Category Premise		5.15	5.21		
Instance Premise		6.77	6.82		

Experiment 7 – Learning

Abstract Structure	Trial Type	Testing Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A 3111_ _	*Classification*	A3111	0.54	0.46	0.69	0.31
A 1311_ _	*Phase 1*	A1311	0.54	0.46	0.69	0.31
A 1131_ _	*Learning Block 1*	A1131	0.58	0.42	0.56	0.44
A 1113_ _		A1113	0.44	0.56	0.56	0.44
A 1111V_		B1333	0.60	0.40	0.56	0.44
A 3113 X_		B3133	0.63	0.38	0.56	0.44
B 1333_ _		B3313	0.54	0.46	0.56	0.44
B 3133_ _		B3331	0.48	0.52	0.50	0.50
B 3313_ _	*Classification*	A3111	0.52	0.48	0.50	0.50
B 3331_ _	*Phase 1*	A1311	0.71	0.29	0.81	0.19
B 3333_Y	*Learning Block 2*	A1131	0.69	0.31	0.63	0.38
B 1331_Z		A1113	0.60	0.40	0.63	0.38
		B1333	0.56	0.44	0.50	0.50
		B3133	0.65	0.35	0.75	0.25
		B3313	0.60	0.40	0.69	0.31
		B3331	0.48	0.52	0.44	0.56
	Classification	A3111	0.56	0.44	0.56	0.44
	Phase 1	A1311	0.71	0.29	0.75	0.25
	Learning Block 3	A1131	0.73	0.27	0.81	0.19
		A1113	0.56	0.44	0.69	0.31
		B1333	0.52	0.48	0.25	0.75
		B3133	0.69	0.31	0.75	0.25
		B3313	0.69	0.31	0.63	0.38
		B3331	0.71	0.29	0.81	0.19
	Classification	A3111	0.67	0.33	0.75	0.25
	Phase 1	A1311	0.65	0.35	0.69	0.31
	Learning Block 4	A1131	0.58	0.42	0.69	0.31
		A1113	0.69	0.31	0.69	0.31
		B1333	0.56	0.44	0.38	0.63
		B3133	0.65	0.35	0.75	0.25
		B3313	0.65	0.35	0.69	0.31
		B3331	0.65	0.35	0.63	0.38
	Classification	A3111	0.67	0.33	0.81	0.19
	Phase 1	A1311	0.69	0.31	0.75	0.25
	Learning Block 5	A1131	0.73	0.27	0.69	0.31
		A1113	0.63	0.38	0.50	0.50
		B1333	0.69	0.31	0.69	0.31
		B3133	0.56	0.44	0.63	0.38
		B3313	0.63	0.38	0.69	0.31
		B3331	0.54	0.46	0.69	0.31
	Classification	A3111	0.58	0.42	0.56	0.44
	Phase 1	A1311	0.71	0.29	0.81	0.19
	Learning Block 6	A1131	0.69	0.31	0.88	0.13
		A1113	0.60	0.40	0.56	0.44
		B1333	0.63	0.38	0.63	0.38
		B3133	0.67	0.33	0.63	0.38
		B3313	0.65	0.35	0.69	0.31
		B3331	0.69	0.31	0.75	0.25
	Classification	A3111	0.75	0.25	0.75	0.25
	Phase 1	A1311	0.65	0.35	0.69	0.31
	Learning Block 7	A1131	0.71	0.29	0.81	0.19
		A1113	0.65	0.35	0.69	0.31
		B1333	0.56	0.44	0.69	0.31
		B3133	0.60	0.40	0.50	0.50
		B3313	0.69	0.31	0.75	0.25
		B3331	0.67	0.33	0.81	0.19
	Classification	A3111	0.75	0.25	0.88	0.13
	Phase 1	A1311	0.79	0.21	0.88	0.13
	Learning Block 8	A1131	0.65	0.35	0.69	0.31
		A1113	0.60	0.40	0.56	0.44
		B1333	0.50	0.50	0.63	0.38
		B3133	0.77	0.23	0.69	0.31
		B3313	0.79	0.21	0.94	0.06
		B3331	0.69	0.31	0.75	0.25
	Classification	A3111	0.73	0.27	0.88	0.13
	Phase 1	A1311	0.71	0.29	0.88	0.13
	Learning Block 9	A1131	0.69	0.31	0.75	0.25
		A1113	0.71	0.29	0.75	0.25
		B1333	0.67	0.33	0.88	0.13
		B3133	0.67	0.33	0.81	0.19
		B3313	0.77	0.23	0.81	0.19
		B3331	0.63	0.38	0.69	0.31

Continued

Classification Phase 1 Learning Block 10	A3111	0.58	0.42	0.69	0.31
	A1311	0.73	0.27	0.88	0.13
	A1131	0.71	0.29	0.88	0.13
	A1113	0.71	0.29	0.75	0.25
	B1333	0.65	0.35	0.88	0.13
	B3133	0.79	0.21	0.94	0.06
	B3313	0.69	0.31	0.94	0.06
	B3331	0.56	0.44	0.69	0.31
Classification Phase 1 Learning Block 11	A3111	0.71	0.29	0.88	0.13
	A1311	0.71	0.29	0.88	0.13
	A1131	0.75	0.25	0.94	0.06
	A1113	0.60	0.40	0.75	0.25
	B1333	0.54	0.46	0.69	0.31
	B3133	0.67	0.33	0.94	0.06
	B3313	0.69	0.31	0.75	0.25
	B3331	0.71	0.29	0.81	0.19
Classification Phase 1 Learning Block 12	A3111	0.63	0.38	0.75	0.25
	A1311	0.73	0.27	0.94	0.06
	A1131	0.73	0.27	0.81	0.19
	A1113	0.69	0.31	0.81	0.19
	B1333	0.65	0.35	0.81	0.19
	B3133	0.75	0.25	0.88	0.13
	B3313	0.73	0.27	0.94	0.06
	B3331	0.67	0.33	0.81	0.19
Classification Phase 1 Learning Block 13	A3111	0.63	0.38	0.94	0.06
	A1311	0.77	0.23	0.88	0.13
	A1131	0.79	0.21	1.00	0.00
	A1113	0.75	0.25	0.88	0.13
	B1333	0.65	0.35	0.69	0.31
	B3133	0.67	0.33	0.81	0.19
	B3313	0.85	0.15	0.94	0.06
	B3331	0.67	0.33	0.75	0.25
Classification Phase 1 Learning Block 14	A3111	0.71	0.29	0.81	0.19
	A1311	0.71	0.29	0.94	0.06
	A1131	0.85	0.15	1.00	0.00
	A1113	0.63	0.38	0.81	0.19
	B1333	0.69	0.31	0.88	0.13
	B3133	0.60	0.40	0.81	0.19
	B3313	0.79	0.21	0.88	0.13
	B3331	0.71	0.29	0.88	0.13
Classification Phase 1 Learning Block 15	A3111	0.69	0.31	0.88	0.13
	A1311	0.71	0.29	0.88	0.13
	A1131	0.77	0.23	1.00	0.00
	A1113	0.79	0.21	0.88	0.13
	B1333	0.71	0.29	0.94	0.06
	B3133	0.77	0.23	0.94	0.06
	B3313	0.75	0.25	0.94	0.06
	B3331	0.81	0.19	0.75	0.25
Classification Phase 1 Learning Block 16	A3111	0.71	0.29	0.88	0.13
	A1311	0.73	0.27	0.75	0.25
	A1131	0.79	0.21	0.81	0.19
	A1113	0.67	0.33	0.75	0.25
	B1333	0.69	0.31	0.81	0.19
	B3133	0.77	0.23	0.88	0.13
	B3313	0.71	0.29	0.81	0.19
	B3331	0.71	0.29	0.88	0.13
Classification Phase 1 Learning Block 17	A3111	0.69	0.31	0.75	0.25
	A1311	0.77	0.23	1.00	0.00
	A1131	0.85	0.15	1.00	0.00
	A1113	0.81	0.19	0.88	0.13
	B1333	0.71	0.29	0.88	0.13
	B3133	0.77	0.23	1.00	0.00
	B3313	0.73	0.27	0.88	0.13
	B3331	0.71	0.29	0.88	0.13
Classification Phase 1 Learning Block 18	A3111	0.79	0.21	0.81	0.19
	A1311	0.75	0.25	0.94	0.06
	A1131	0.81	0.19	0.94	0.06
	A1113	0.71	0.29	0.94	0.06
	B1333	0.65	0.35	0.75	0.25
	B3133	0.73	0.27	0.94	0.06
	B3313	0.77	0.23	0.88	0.13
	B3331	0.71	0.29	0.88	0.13
Classification Phase 1 Learning Block 19	A3111	0.73	0.27	0.88	0.13
	A1311	0.83	0.17	1.00	0.00
	A1131	0.77	0.23	0.94	0.06
	A1113	0.65	0.35	0.94	0.06
	B1333	0.71	0.29	0.88	0.13
	B3133	0.75	0.25	0.94	0.06
	B3313	0.81	0.19	0.94	0.06
	B3331	0.73	0.27	1.00	0.00
Classification Phase 1 Learning Block 20	A3111	0.69	0.31	0.94	0.06
	A1311	0.73	0.27	0.88	0.13
	A1131	0.79	0.21	1.00	0.00
	A1113	0.73	0.27	0.94	0.06
	B1333	0.71	0.29	0.88	0.13
	B3133	0.67	0.33	0.94	0.06
	B3313	0.79	0.21	0.94	0.06
	B3331	0.71	0.29	0.81	0.19

Continued

Classification Phase 2 Learning Block 1	A111 1V_	0.85	0.15	0.88	0.13
	A311 3X_	0.56	0.44	0.50	0.50
	B3333_ Y	0.79	0.21	0.88	0.13
	B1331_ Z	0.40	0.60	0.31	0.69
Classification Phase 2 Learning Block 2	A111 1V_	0.90	0.10	0.94	0.06
	A311 3X_	0.71	0.29	0.88	0.13
	B3333_ Y	0.83	0.17	0.94	0.06
	B1331_ Z	0.75	0.25	0.88	0.13
Classification Phase 2 Learning Block 3	A111 1V_	0.90	0.10	0.93	0.07
	A311 3X_	0.83	0.17	0.93	0.07
	B3333_ Y	0.85	0.15	1.00	0.00
	B1331_ Z	0.80	0.20	0.87	0.13
Classification Phase 2 Learning Block 4	A111 1V_	0.90	0.10	0.93	0.07
	A311 3X_	0.80	0.20	0.93	0.07
	B3333_ Y	0.94	0.06	0.93	0.07
	B1331_ Z	0.80	0.20	0.93	0.07
Classification Phase 2 Learning Block 5	A111 1V_	0.92	0.08	0.94	0.06
	A311 3X_	0.90	0.10	0.94	0.06
	B3333_ Y	0.94	0.06	1.00	0.00
	B1331_ Z	0.85	0.15	1.00	0.00
Classification Phase 2 Learning Block 6	A111 1V_	0.92	0.08	0.94	0.06
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.94	0.06	0.94	0.06
	B1331_ Z	0.88	0.13	0.88	0.13
Classification Phase 2 Learning Block 7	A111 1V_	0.96	0.04	1.00	0.00
	A311 3X_	0.92	0.08	0.94	0.06
	B3333_ Y	0.92	0.08	0.94	0.06
	B1331_ Z	0.90	0.10	0.94	0.06
Classification Phase 2 Learning Block 8	A111 1V_	0.98	0.02	1.00	0.00
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.92	0.08	1.00	0.00
	B1331_ Z	0.85	0.15	0.94	0.06
Classification Phase 2 Learning Block 9	A111 1V_	0.98	0.02	0.94	0.06
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_ Z	0.90	0.10	0.94	0.06
Classification Phase 2 Learning Block 10	A111 1V_	0.98	0.02	1.00	0.00
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.98	0.02	1.00	0.00
	B1331_ Z	0.92	0.08	1.00	0.00
Classification Phase 2 Learning Block 11	A111 1V_	0.96	0.04	0.94	0.06
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.94	0.06	1.00	0.00
	B1331_ Z	0.92	0.08	1.00	0.00
Classification Phase 2 Learning Block 12	A111 1V_	0.96	0.04	0.94	0.06
	A311 3X_	0.92	0.08	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_ Z	0.94	0.06	1.00	0.00
Classification Phase 2 Learning Block 13	A111 1V_	0.98	0.02	0.94	0.06
	A311 3X_	0.90	0.10	0.94	0.06
	B3333_ Y	0.98	0.02	1.00	0.00
	B1331_ Z	0.94	0.06	1.00	0.00
Classification Phase 2 Learning Block 14	A111 1V_	0.94	0.06	1.00	0.00
	A311 3X_	0.94	0.06	1.00	0.00
	B3333_ Y	0.98	0.02	1.00	0.00
	B1331_ Z	0.94	0.06	1.00	0.00
Classification Phase 2 Learning Block 15	A111 1V_	0.96	0.04	1.00	0.00
	A311 3X_	0.92	0.08	1.00	0.00
	B3333_ Y	0.92	0.08	1.00	0.00
	B1331_ Z	0.90	0.10	1.00	0.00
Classification Phase 2 Learning Block 16	A111 1V_	0.96	0.04	1.00	0.00
	A311 3X_	0.90	0.10	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_ Z	0.88	0.13	0.94	0.06
Classification Phase 2 Learning Block 17	A111 1V_	1.00	0.00	1.00	0.00
	A311 3X_	0.94	0.06	1.00	0.00
	B3333_ Y	0.94	0.06	1.00	0.00
	B1331_ Z	0.90	0.10	1.00	0.00
Classification Phase 2 Learning Block 18	A111 1V_	0.96	0.04	1.00	0.00
	A311 3X_	0.96	0.04	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_ Z	0.94	0.06	1.00	0.00
Classification Phase 2 Learning Block 19	A111 1V_	0.96	0.04	1.00	0.00
	A311 3X_	0.92	0.08	1.00	0.00
	B3333_ Y	0.94	0.06	1.00	0.00
	B1331_ Z	0.98	0.02	1.00	0.00
Classification Phase 2 Learning Block 20	A111 1V_	0.94	0.06	1.00	0.00
	A311 3X_	0.94	0.06	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_ Z	0.92	0.08	1.00	0.00

Experiment 7 – Testing

Trial Type	Testing Trials	No Criterion		Criterion	
		Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
Ordinary Premise Typicality	A3111?_	0.56	0.44	0.50	0.50
	A1113?_	0.50	0.50	0.56	0.44
	B1333_?	0.42	0.58	0.31	0.69
	B3331_?	0.48	0.52	0.56	0.44
Hidden Feature Inference Block 1	A1111V_	0.58	0.42	0.44	0.56
	A3113X_	0.50	0.50	0.44	0.56
	B3333_Y	0.48	0.52	0.44	0.56
	B1331_Z	0.52	0.48	0.50	0.50
Generalized Premise Typicality	A1133?_	0.52	0.48	0.56	0.44
	B3311_?	0.40	0.60	0.31	0.69
	A1313?_	0.67	0.33	0.63	0.38
	B3131_?	0.40	0.60	0.31	0.69
Hidden Feature Inference Block 2	A1111V_	0.50	0.50	0.44	0.56
	A3113X_	0.40	0.60	0.50	0.50
	B3333_Y	0.46	0.54	0.50	0.50
	B1331_Z	0.48	0.52	0.50	0.50
Classification with Hidden Features	A3111_ _	0.79	0.21	0.94	0.06
	A1311_ _	0.69	0.31	1.00	0.00
	A1131_ _	0.81	0.19	0.94	0.06
	A1113_ _	0.79	0.21	0.88	0.13
	A1111V _	0.96	0.04	1.00	0.00
	A3113X _	0.98	0.02	1.00	0.00
	B1333_ _	0.71	0.29	1.00	0.00
	B3133_ _	0.69	0.31	0.94	0.06
	B3313_ _	0.81	0.19	1.00	0.00
	B3331_ _	0.81	0.19	1.00	0.00
	B3333_ Y	0.96	0.04	1.00	0.00
	B1331_Z	0.94	0.06	1.00	0.00
Premise Conclusion Similarity	A1311?_	0.56	0.44	0.56	0.44
	A1131?_	0.60	0.40	0.50	0.50
	B3133_?	0.42	0.58	0.50	0.50
	B3313_?	0.50	0.50	0.50	0.50
Hidden Feature Inference Block 3	A1111V_	0.52	0.48	0.44	0.56
	A3113X_	0.48	0.52	0.38	0.63
	B3333_Y	0.52	0.48	0.63	0.38
	B1331_Z	0.50	0.50	0.44	0.56
Hidden Feature Inference Block 4	A1111V_	0.58	0.42	0.38	0.63
	A3113X_	0.48	0.52	0.50	0.50
	B3333_Y	0.58	0.42	0.69	0.31
	B1331_Z	0.42	0.58	0.31	0.69
Hidden Feature Inference Block 5	A1111V_	0.58	0.42	0.69	0.31
	A3113X_	0.54	0.46	0.63	0.38
	B3333_Y	0.48	0.52	0.50	0.50
	B1331_Z	0.46	0.54	0.69	0.31
Ambiguous Exception Feature Inference	A?111_ _	0.67	0.33	0.75	0.25
	B?333_ _	0.65	0.35	0.75	0.25
	A111?_ _	0.60	0.40	0.69	0.31
	B333?_ _	0.69	0.31	0.88	0.13
	A3113X_	0.44	0.56	0.44	0.56
	A3113X_	0.44	0.56	0.31	0.69
	B1331_Y	0.50	0.50	0.63	0.38
	B1331_Y	0.31	0.69	0.25	0.75
Ordinary Premise Typicality	A3111?_	0.58	0.42	0.50	0.50
	A1113?_	0.52	0.48	0.44	0.56
	B1333_?	0.48	0.52	0.56	0.44
	B3331_?	0.48	0.52	0.69	0.31
Hidden Feature Inference Block 6	A1111V_	0.56	0.44	0.44	0.56
	A3113X_	0.38	0.63	0.38	0.63
	B3333_Y	0.56	0.44	0.56	0.44
	B1331_Z	0.48	0.52	0.44	0.56

Continued

Generalized Premise Typicality	A1133?_	0.56	0.44	0.44	0.56
	B3311_?	0.56	0.44	0.44	0.56
	A1313?_	0.50	0.50	0.56	0.44
	B3131_?	0.48	0.52	0.44	0.56
Hidden Feature Inference Block 7	A1111V_	0.54	0.46	0.50	0.50
	A3113X_	0.50	0.50	0.63	0.38
	B3333_Y	0.52	0.48	0.56	0.44
	B1331_Z	0.44	0.56	0.44	0.56
Classification	A3111_ _	0.75	0.25	0.88	0.13
	A1311_ _	0.77	0.23	0.94	0.06
	A1131_ _	0.83	0.17	0.88	0.13
	A1113_ _	0.79	0.21	0.94	0.06
	A1111_ _	0.88	0.13	1.00	0.00
	A3113_ _	0.44	0.56	0.31	0.69
	B1333_ _	0.79	0.21	1.00	0.00
	B3133_ _	0.69	0.31	0.94	0.06
	B3313_ _	0.85	0.15	0.88	0.13
	B3331_ _	0.83	0.17	1.00	0.00
	B3333_ _	0.92	0.08	1.00	0.00
	B1331_ _	0.50	0.50	0.38	0.63
Label vs Feature	A3?33_Y	0.54	0.46	0.69	0.31
	B1?11_V	0.63	0.38	0.75	0.25
Continuous Generalization	A2112?_	0.56	0.44	0.38	0.63
	B1221_?	0.38	0.63	0.31	0.69
	A0110?_	0.46	0.54	0.44	0.56
	B4114_?	0.40	0.60	0.56	0.44
	A1221?_	0.52	0.48	0.56	0.44
	B3223_?	0.46	0.54	0.44	0.56
	A1001?_	0.52	0.48	0.31	0.69
	B3443_?	0.50	0.50	0.56	0.44
Blank Feature Inference	A_ _ _ _?_	0.63	0.38	0.44	0.56
	B_ _ _ _ _?	0.65	0.35	0.69	0.31
Label vs Hidden Features	A_?_ _ _Y	0.56	0.44	0.56	0.44
	A_ _?_ _Z	0.65	0.35	0.75	0.25
	B_?_ _ W_	0.69	0.31	0.50	0.50
	B_ _?_X_	0.63	0.38	0.81	0.19
Generalized Classification with Hidden Features	?1133V	0.73	0.27	0.69	0.31
	?3311X_	0.35	0.65	0.31	0.69
	?1313_Y	0.27	0.73	0.31	0.69
	?3131_Z	0.73	0.27	0.94	0.06
Generalized Classification	?1133_ _	0.50	0.50	0.44	0.56
	?3311_ _	0.58	0.42	0.50	0.50
	?1313_ _	0.44	0.56	0.38	0.63
	?3131_ _	0.54	0.46	0.75	0.25
Premise Diversity	A2212?_	0.50	0.50	0.38	0.63
	B2232_?	0.60	0.40	0.63	0.38
The Inclusion Fallacy	A_ _ _ _?_	0.48	0.52	0.38	0.63
	A3003?_	0.52	0.48	0.50	0.50
	B_ _ _ _ _?	0.50	0.50	0.44	0.56
	B1441_?	0.48	0.52	0.31	0.69
Typical Premise		5.35	5.31		
Atypical Premise		4.85	5.38		
Typical Conclusion		5.29	5.13		
Atypical Conclusion		4.75	4.81		
More Diverse Premise		5.88	5.31		
Less Diverse Premise		6.29	5.63		
Category Conclusion		6.02	6.06		
Instance Conclusion		4.52	4.56		
Category Premise		6.25	7.00		
Instance Premise		7.56	7.44		

Experiment 8 – Learning

Abstract Structure	Trial Type	Trials	No Criterion		Criterion	
			Correct Typical Label	Incorrect Atypical HFs	Correct Typical Label	Incorrect Atypical HFs
A3111_ _ A1311_ _ A1131_ _ A1113_ _ A1111V_ A3113 X_ B1333_ _ B3133_ _ B3313_ _ B3331_ _ B3333_Y B1331_Z	*Learning Phase 1 Block 1*	A3111	0.57	0.43	0.52	0.48
		A1311	0.54	0.46	0.69	0.31
		A1131	0.55	0.45	0.52	0.48
		A1113	0.56	0.44	0.62	0.38
		B1333	0.45	0.55	0.45	0.55
		B3133	0.51	0.49	0.38	0.62
		B3313	0.48	0.52	0.55	0.45
		B3331	0.52	0.48	0.59	0.41
	Learning Phase 1 Block 2	A3111	0.60	0.40	0.66	0.34
		A1311	0.56	0.44	0.62	0.38
		A1131	0.55	0.45	0.52	0.48
		A1113	0.59	0.41	0.66	0.34
		B1333	0.53	0.47	0.59	0.41
		B3133	0.59	0.41	0.52	0.48
		B3313	0.63	0.38	0.59	0.41
		B3331	0.60	0.40	0.66	0.34
	Learning Phase 1 Block 3	A3111	0.59	0.41	0.69	0.31
		A1311	0.55	0.45	0.66	0.34
		A1131	0.54	0.46	0.55	0.45
		A1113	0.58	0.42	0.66	0.34
		B1333	0.59	0.41	0.62	0.38
		B3133	0.54	0.46	0.66	0.34
		B3313	0.57	0.43	0.66	0.34
		B3331	0.56	0.44	0.55	0.45
	Learning Phase 1 Block 4	A3111	0.70	0.30	0.90	0.10
		A1311	0.62	0.38	0.62	0.38
		A1131	0.58	0.42	0.52	0.48
		A1113	0.65	0.35	0.69	0.31
		B1333	0.57	0.43	0.59	0.41
		B3133	0.61	0.39	0.69	0.31
		B3313	0.58	0.42	0.66	0.34
		B3331	0.65	0.35	0.72	0.28
	Learning Phase 1 Block 5	A3111	0.68	0.32	0.79	0.21
		A1311	0.66	0.34	0.76	0.24
		A1131	0.56	0.44	0.52	0.48
		A1113	0.66	0.34	0.72	0.28
		B1333	0.64	0.36	0.79	0.21
		B3133	0.65	0.35	0.69	0.31
		B3313	0.54	0.46	0.52	0.48
		B3331	0.67	0.33	0.59	0.41

Continued

Learning Phase 2 Block 1	A3111__	0.69	0.31	0.76	0.24
	A1311__	0.70	0.30	0.66	0.34
	A1131__	0.59	0.41	0.48	0.52
	A1113__	0.67	0.33	0.72	0.28
	A1111__	0.87	0.13	0.86	0.14
	A3113__	0.48	0.52	0.45	0.55
	B1333__	0.74	0.26	0.76	0.24
	B3133__	0.66	0.34	0.69	0.31
	B3313__	0.65	0.35	0.55	0.45
	B3331__	0.66	0.34	0.83	0.17
	B3333__	0.77	0.23	0.83	0.17
	B1331__	0.55	0.45	0.62	0.38
Learning Phase 2 Block 2	A3111__	0.73	0.27	0.83	0.17
	A1311__	0.59	0.41	0.62	0.38
	A1131__	0.59	0.41	0.62	0.38
	A1113__	0.66	0.34	0.76	0.24
	A1111__	0.80	0.20	0.93	0.07
	A3113__	0.55	0.45	0.72	0.28
	B1333__	0.79	0.21	0.83	0.17
	B3133__	0.58	0.42	0.76	0.24
	B3313__	0.52	0.48	0.66	0.34
	B3331__	0.77	0.23	0.83	0.17
	B3333__	0.82	0.18	1.00	0.00
	B1331__	0.58	0.42	0.66	0.34
Learning Phase 2 Block 3	A3111__	0.74	0.26	0.90	0.10
	A1311__	0.55	0.45	0.59	0.41
	A1131__	0.61	0.39	0.45	0.55
	A1113__	0.71	0.29	0.79	0.21
	A1111__	0.82	0.18	0.93	0.07
	A3113__	0.59	0.41	0.66	0.34
	B1333__	0.76	0.24	0.69	0.31
	B3133__	0.55	0.45	0.45	0.55
	B3313__	0.64	0.36	0.72	0.28
	B3331__	0.71	0.29	0.86	0.14
	B3333__	0.86	0.14	0.97	0.03
	B1331__	0.70	0.30	0.90	0.10
Learning Phase 2 Block 4	A3111__	0.81	0.19	0.97	0.03
	A1311__	0.57	0.43	0.59	0.41
	A1131__	0.59	0.41	0.66	0.34
	A1113__	0.77	0.23	0.93	0.07
	A1111__	0.84	0.16	0.97	0.03
	A3113__	0.65	0.35	0.83	0.17
	B1333__	0.75	0.25	0.83	0.17
	B3133__	0.63	0.37	0.79	0.21
	B3313__	0.58	0.42	0.62	0.38
	B3331__	0.73	0.27	0.86	0.14
	B3333__	0.85	0.15	0.97	0.03
	B1331__	0.78	0.22	0.86	0.14
Learning Phase 2 Block 5	A3111__	0.76	0.24	0.90	0.10
	A1311__	0.55	0.45	0.69	0.31
	A1131__	0.60	0.40	0.69	0.31
	A1113__	0.80	0.20	0.93	0.07
	A1111__	0.88	0.13	0.97	0.03
	A3113__	0.67	0.33	0.79	0.21
	B1333__	0.80	0.20	0.97	0.03
	B3133__	0.65	0.35	0.72	0.28
	B3313__	0.58	0.42	0.62	0.38
	B3331__	0.75	0.25	0.90	0.10
	B3333__	0.86	0.14	0.90	0.10
	B1331__	0.70	0.30	0.86	0.14

Continued

Learning Phase 3 Block 1	A3111_ _	0.79	0.21	0.90	0.10
	A1311_ _	0.68	0.32	0.86	0.14
	A1131_ _	0.66	0.34	0.72	0.28
	A1113_ _	0.77	0.23	0.97	0.03
	A1111V _	0.75	0.25	0.93	0.07
	A3113X _	0.67	0.33	0.86	0.14
	B1333_ _	0.79	0.21	0.93	0.07
	B3133_ _	0.60	0.40	0.86	0.14
	B3313_ _	0.58	0.42	0.72	0.28
	B3331_ _	0.77	0.23	0.86	0.14
	B3333Y _	0.77	0.23	0.86	0.14
	B1331Z _	0.74	0.26	0.86	0.14
	A1111V _	0.65	0.35	0.66	0.34
	A3113X_	0.41	0.59	0.34	0.66
	B3333_ Y	0.20	0.80	0.07	0.93
	B1331_ Z	0.55	0.45	0.59	0.41
Learning Phase 3 Block 2	A3111_ _	0.80	0.20	0.86	0.14
	A1311_ _	0.62	0.38	0.83	0.17
	A1131_ _	0.54	0.46	0.59	0.41
	A1113_ _	0.78	0.22	0.90	0.10
	A1111V _	0.88	0.12	0.97	0.03
	A3113X _	0.70	0.30	0.93	0.07
	B1333_ _	0.80	0.20	0.83	0.17
	B3133_ _	0.60	0.40	0.62	0.38
	B3313_ _	0.55	0.45	0.66	0.34
	B3331_ _	0.78	0.22	0.93	0.07
	B3333Y _	0.82	0.18	0.97	0.03
	B1331Z _	0.73	0.27	0.86	0.14
	A1111V _	0.54	0.46	0.52	0.48
	A3113X_	0.45	0.55	0.48	0.52
	B3333_ Y	0.45	0.55	0.38	0.62
	B1331_ Z	0.62	0.38	0.69	0.31
Learning Phase 3 Block 3	A3111_ _	0.80	0.20	0.93	0.07
	A1311_ _	0.66	0.34	0.90	0.10
	A1131_ _	0.59	0.41	0.66	0.34
	A1113_ _	0.76	0.24	0.90	0.10
	A1111V _	0.85	0.15	0.97	0.03
	A3113X _	0.68	0.32	0.83	0.17
	B1333_ _	0.21	0.79	0.07	0.93
	B3133_ _	0.45	0.55	0.31	0.69
	B3313_ _	0.41	0.59	0.48	0.52
	B3331_ _	0.76	0.24	0.88	0.12
	B3333Y _	0.87	0.13	1.00	0.00
	B1331Z _	0.25	0.75	0.10	0.90
	A1111V _	0.67	0.33	0.76	0.24
	A3113X_	0.45	0.55	0.48	0.52
	B3333_ Y	0.46	0.54	0.48	0.52
	B1331_ Z	0.63	0.38	0.76	0.24
Learning Phase 3 Block 4	A3111_ _	0.83	0.17	0.97	0.03
	A1311_ _	0.62	0.38	0.86	0.14
	A1131_ _	0.66	0.34	0.76	0.24
	A1113_ _	0.79	0.21	0.93	0.07
	A1111V _	0.88	0.13	0.97	0.03
	A3113X _	0.70	0.30	0.86	0.14
	B1333_ _	0.80	0.20	0.93	0.07
	B3133_ _	0.63	0.37	0.76	0.24
	B3313_ _	0.63	0.38	0.76	0.24
	B3331_ _	0.74	0.26	1.00	0.00
	B3333Y _	0.86	0.14	1.00	0.00
	B1331Z _	0.77	0.23	0.93	0.07
	A1111V _	0.73	0.27	0.90	0.10
	A3113X_	0.53	0.47	0.62	0.38
	B3333_ Y	0.45	0.55	0.55	0.45
	B1331_ Z	0.64	0.36	0.86	0.14
Learning Phase 3 Block 5	A3111_ _	0.82	0.18	1.00	0.00
	A1311_ _	0.59	0.41	0.83	0.17
	A1131_ _	0.59	0.41	0.79	0.21
	A1113_ _	0.78	0.22	0.93	0.07
	A1111V _	0.85	0.15	0.97	0.03
	A3113X _	0.73	0.27	0.93	0.07
	B1333_ _	0.73	0.27	0.93	0.07
	B3133_ _	0.59	0.41	0.59	0.41
	B3313_ _	0.58	0.42	0.66	0.34
	B3331_ _	0.73	0.27	0.86	0.14
	B3333Y _	0.88	0.13	1.00	0.00
	B1331Z _	0.73	0.27	0.86	0.14
	A1111V _	0.71	0.29	0.76	0.24
	A3113X_	0.56	0.44	0.59	0.41
	B3333_ Y	0.51	0.49	0.66	0.34
	B1331_ Z	0.68	0.32	0.83	0.17

Experiment 8 – Testing

Trial Type	Testing Trials	No Criterion Correct Typical Label	No Criterion Incorrect Atypical HFs	Criterion Correct Typical Label	Criterion Incorrect Atypical HFs
Ordinary Premise Typicality Block 1	A3111?_	0.54	0.46	0.49	0.51
	A1113?_	0.49	0.51	0.47	0.54
	B1333_?	0.48	0.52	0.54	0.47
	B3331_?	0.53	0.47	0.58	0.42
Hidden Feature Inference Block 1	A1111V_	0.70	0.31	0.84	0.16
	A3113X_	0.62	0.38	0.81	0.19
	B3333_Y	0.63	0.38	0.86	0.14
	B1331_Z	0.68	0.32	0.86	0.14
Generalized Premise Typicality Block 1	A1133?_	0.52	0.48	0.47	0.54
	B3311_?	0.50	0.50	0.51	0.49
	A1313?_	0.60	0.40	0.58	0.42
	B3131_?	0.54	0.46	0.54	0.47
Hidden Feature Inference Block 2	A1111V_	0.73	0.27	0.86	0.14
	A3113X_	0.36	0.64	0.12	0.88
	B3333_Y	0.69	0.31	0.93	0.07
	B1331_Z	0.34	0.66	0.05	0.95
Classification with Hidden Features	A3111_ _	0.77	0.23	0.88	0.12
	A1311_ _	0.69	0.31	0.74	0.26
	A1131_ _	0.68	0.32	0.67	0.33
	A1113_ _	0.84	0.16	0.95	0.05
	A1111V_	0.91	0.09	1.00	0.00
	A3113X_	0.74	0.26	0.93	0.07
	B1333_ _	0.81	0.19	0.93	0.07
	B3133_ _	0.63	0.37	0.79	0.21
	B3313_ _	0.68	0.32	0.70	0.30
	B3331_ _	0.76	0.24	0.86	0.14
	B3333_Y	0.89	0.11	0.98	0.02
	B1331_Z	0.79	0.21	1.00	0.00
Premise Conclusion Similarity	A1311?_	0.65	0.35	0.81	0.19
	A1131?_	0.66	0.34	0.88	0.12
	B3133_?	0.63	0.37	0.88	0.12
	B3313_?	0.66	0.34	0.91	0.09
Hidden Feature Inference Block 3	A1111V_	0.66	0.34	0.88	0.12
	A3113X_	0.59	0.41	0.81	0.19
	B3333_Y	0.66	0.34	0.88	0.12
	B1331_Z	0.68	0.32	0.84	0.16
Hidden Feature Inference Block 4	A1111V_	0.69	0.31	0.88	0.12
	A3113X_	0.64	0.36	0.84	0.16
	B3333_Y	0.65	0.35	0.86	0.14
	B1331_Z	0.70	0.30	0.95	0.05
Hidden Feature Inference Block 5	A1111V_	0.66	0.34	0.86	0.14
	A3113X_	0.64	0.36	0.77	0.23
	B3333_Y	0.66	0.34	0.91	0.09
	B1331_Z	0.69	0.31	0.86	0.14
Ambiguous	A?111_ _	0.61	0.39	0.65	0.35
	B?333_ _	0.62	0.38	0.58	0.42
	A111?_ _	0.52	0.48	0.63	0.37
	B333?_ _	0.60	0.40	0.72	0.28
Exception Feature Inference	A3113X_	0.49	0.51	0.51	0.49
	A3113X_	0.47	0.53	0.53	0.47
	B1331_Y	0.53	0.47	0.65	0.35
	B1331_Y	0.51	0.49	0.58	0.42

Continued

Ordinary Premise Typicality Block 2	A3111?_	0.59	0.41	0.58	0.42
	A1113?_	0.58	0.42	0.47	0.54
	B1333_?	0.56	0.45	0.49	0.51
	B3331_?	0.53	0.47	0.49	0.51
Hidden Feature Inference Block 6	A1111V_	0.65	0.35	0.88	0.12
	A3113X_	0.61	0.39	0.81	0.19
	B3333_Y	0.59	0.41	0.86	0.14
	B1331_Z	0.70	0.31	0.93	0.07
Generalized Premise Typicality Block 2	A1133?_	0.49	0.51	0.44	0.56
	B3311_?	0.52	0.48	0.44	0.56
	A1313?_	0.57	0.43	0.51	0.49
	B3131_?	0.55	0.45	0.54	0.47
Hidden Feature Inference Block 7	A1111V_	0.70	0.31	0.79	0.21
	A3113X_	0.59	0.41	0.77	0.23
	B3333_Y	0.72	0.28	0.93	0.07
	B1331_Z	0.70	0.31	0.91	0.09
Classification	A3111_ _	0.75	0.25	0.91	0.09
	A1311_ _	0.71	0.30	0.86	0.14
	A1131_ _	0.68	0.32	0.67	0.33
	A1113_ _	0.71	0.30	0.86	0.14
	A1111_ _	0.86	0.14	0.93	0.07
	A3113_ _	0.66	0.34	0.70	0.30
	B1333_ _	0.77	0.23	0.95	0.05
	B3133_ _	0.64	0.36	0.77	0.23
	B3313_ _	0.65	0.35	0.72	0.28
	B3331_ _	0.82	0.18	0.93	0.07
	B3333_ _	0.85	0.15	0.95	0.05
	B1331_ _	0.66	0.34	0.77	0.23
Label vs Feature	A3?33_Y	0.59	0.41	0.63	0.37
	B1?11_V	0.58	0.42	0.54	0.47
Blank Feature Inference	A_ _ _ _?_	0.56	0.45	0.61	0.40
	B_ _ _ _ _?_	0.49	0.50	0.47	0.54
Label vs Hidden Feature	A_?_ _ _Y	0.56	0.44	0.56	0.44
	A_ _?_ _Z	0.68	0.32	0.61	0.40
	B_?_ _W_	0.65	0.35	0.67	0.33
	B_ _ ?_X_	0.65	0.35	0.70	0.30
Generalized Classification Block 1	?1133_ _	0.45	0.56	0.33	0.67
	?3311_ _	0.60	0.40	0.72	0.28
	?1313_ _	0.56	0.45	0.49	0.51
	?3131_ _	0.50	0.50	0.58	0.42
Generalized Classification Block 2	?1133_ _	0.46	0.54	0.33	0.67
	?3311_ _	0.56	0.45	0.65	0.35
	?1313_ _	0.48	0.52	0.44	0.56
	?3131_ _	0.46	0.54	0.56	0.44
Typical Premise		5.7	5.8		
Atypical Premise		4.9	4.7		
Typical Conclusion		5.9	5.9		
Atypical Conclusion		4.8	4.8		
More Diverse Premise		6.0	5.6		
Less Diverse Premise		6.3	6.7		
Category Conclusion		6.4	6.2		
Instance Conclusion		4.8	4.2		
Category Premise		7.8	7.8		
Category Instance		5.9	6.0		

7.3. Appendix C: Additional testing trials included in Experiments 4-8.

Experiments 4-8 included additional testing trials that were not central to the key assessments of premise typicality. These are described in detail below and were intended to provide additional clarification and constraints for models. The results for these trials are presented in Appendix D and the average response proportions across participants are in Appendix B.

Experiments 4-8 included 'ambiguous' testing trials that matched two instances in the category summary, the typical instance and an ordinary instance and these predicted different features as a response. For example, the instance A?111 has the same last three features as both the typical instance A1111 and the ordinary instance A3111. So, based on a match to a single category instance, both 1 and 3 are possible responses, however, a 1 feature is the more typical feature for category A, so a 1 feature response potentially corresponds to a typicality effect.

Exploratory trials in Experiments 4-8 evaluated the relative influence that each part of the stimulus had on responding—category labels, non-hidden features and hidden features—by pitting these against each other. In the 'label vs feature' trials the category label from one category was combined with the typical features of the other category and participants were queried on a missing non-hidden feature. So, the label corresponded to the typical response for one category while the instance features corresponded to the alternative response i.e. the typical features of the other category. Similarly, the 'label vs hidden feature' trials contrasted a feature inference response consistent with the category label (the typical non-hidden feature for that category) to the response consistent with the hidden feature (the atypical non-hidden feature for the category denoted by the category label).

Another common effect in categorical induction is premise diversity in which the conclusion is judged as stronger when the premises of an argument are diverse in their coverage of a category. Experiments 4-6 tested premise diversity by adding a hidden feature to one

additional instance in each category in the category summary (specifically the A1311 and B3133 instances) that was typical and atypical respectively. Therefore, category A had a less diverse set of instances with the typical feature (A1111 and A1311) whereas category B had a more diverse set of instances with the atypical feature (B1331 and B3133). The test instances were A2212 and B2232 which were equally similar to the typical and atypical instances for each category and so there should have been no effect of similarity. These are continuous instances as the 2 value relates to the feature dimensions on a continuum from 0 to 4 (see Figure 53). Additionally, putting the atypical hidden feature on the more diverse set of premises separates off the effect of premise typicality (and the assumed preference for the typical hidden feature) from the effects of premise diversity. This is because a premise diversity effect could be detected as a 'premise atypicality' effect on these trials whereas if the typical hidden feature was attached to the more diverse set, a preference for responding with the typical hidden feature

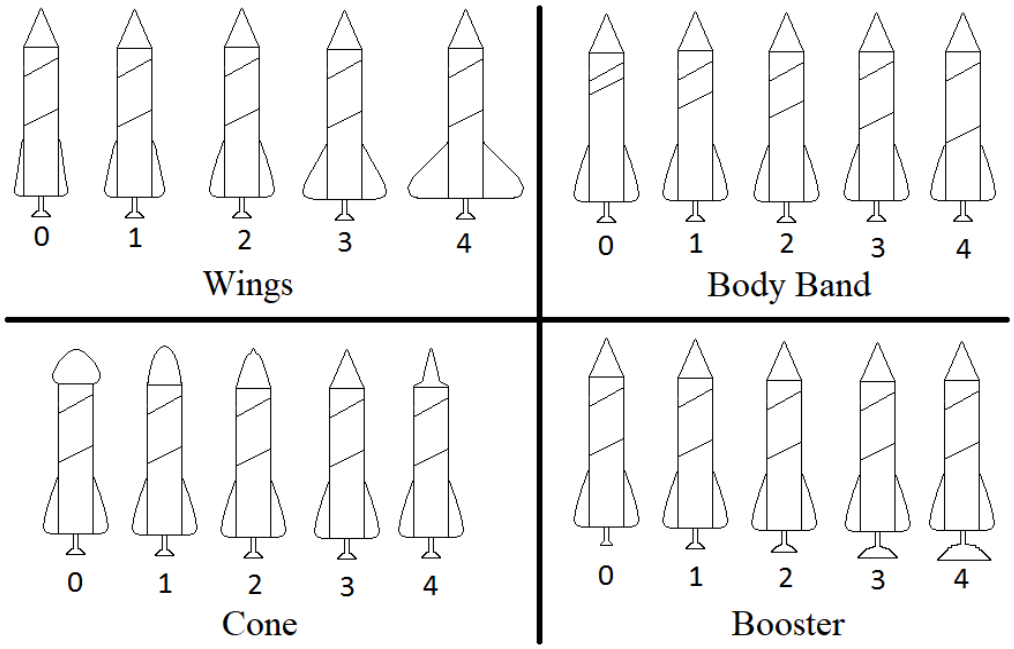

Figure 53. Specification of the continuous dimension values from 0 to 4 for the four non-hidden feature dimensions used in Experiments 4-7. Top left panel = wing continuum, top right panel = body band continuum, bottom left panel = cone continuum, bottom right panel = booster continuum.

might have been seen due to premise typicality, premise diversity or both. So specifically, a diversity effect would correspond to more responding with the atypical feature on the test instance in category B than for the typical feature on the test instance in category A. Note, that the lack of category summary in Experiment 7 meant that these trials were not tests of premise diversity in this experiment though they were included in the testing phase.

A further categorical induction effect is the inclusion fallacy which occurs when a conclusion that covers a category is judged stronger than a conclusion that is a member of that category. Experiments 4-7 attempted to test the inclusion fallacy via a blank feature inference trial (a trial with no feature information present on the screen, only a category label is presented) and a specific category instance for each category. The specific category instances were A3003 and B1441. Continuous instances were used as all non-continuous instances had been used in other testing trials and may have had associations separate from the inclusion fallacy. A3003 and B1441 had not been used in any other testing trials. We originally thought an inclusion fallacy would manifest as higher levels of typical responding on the blank trials than the specific instance trials. However, we've realized after the fact that this is not well founded as the specific instance is also more similar to the atypical instance. Also as discussed below the specific instance was based on the continuous instances which were likely perceptually problematic. As a result, I have not discussed the results of these trials below and have only reported the classic paradigm test results for the inclusion fallacy.

Experiments 4-8 also included generalization classification trials which queried the category label for the four instances that were not present in the category summary/not trained during the training phase (A1133, B3311, A1313, B3131). These trials tested for dimensional rule use.

The stimuli used in Experiments 4-7 were composed of features on continuous dimensions that had five possible values (see Figure 53). For example, the wing dimension on

the continuum varied from a value of 0 which had very narrow wings to the value of 4 which had very wide wings. The wings presented as part of the category summary or learning task had a value of 1 for the narrow wings and a value of 3 for the wide wings (note this is not a direct reference to the abstract category structures as the assignment of physical continuum values to abstract features values was counterbalanced so that '1' abstractly could refer to wide or narrow wings). The wings with the value of 2 was intermediate in size between the wide and narrow wings. These continuums allowed for 'continuous generalization' testing trials and these were used to test similarity. For example, A0110? tested similarity as this item is more similar to the typical than to the atypical instance. The single case, A2112?, was equally similar to the typical and atypical instances, potentially measuring continuous premise typicality. However, as this was only a single instance and was based on the potentially inconsistent continuous stimuli, this was not used as a reliable measure of premise typicality.

Finally, the classic version of some common categorical induction effects were also included alongside the classic version of premise typicality as a further check that these effects could be obtained in the current participant population. These additional tests included: conclusion typicality, premise diversity, the inclusion fallacy and premise specificity.

7.4. Appendix D: Results for subsidiary testing trials described in Appendix C.

Experiment 4

In Experiment 4, the ambiguous testing trials did not show a significant preference for responding with the typical non-hidden features over the atypical, ($t(36) = 1.3$, $p = 0.212$). This indicates no typicality effect with non-hidden features. The label vs feature trials showed marginally significant non-hidden feature consistent responding ($t(36) = 1.7$, $p = 0.095$). That is, participants were responding with the feature typical of the category denoted by the non-hidden features e.g. responding with the 1 feature at test when all the non-hidden features also had a 1 value. The label vs hidden feature trials had no significant preference in responding based on the label over the hidden features ($t(36) = 0.4$, $p = 0.657$). Overall, these trials show only a weak preference for the non-hidden features in responding. The generalized classification testing trials showed that 32% of participants responded in a way that was perfectly consistent with a rule. There were four patterns of responding that matched to each of the four possible unidimensional rules. Any participant who matched the pattern of responses predicted by any of the four rules over all four classification generalization trials was classed as responding in a rule consistent way. For example, the pattern matching to a unidimensional rule on dimension one would have errors on instances 2 and 4 but not on instances 1 and 3. This percentage suggests that roughly a third of participants were responding consistent with rule use.

The continuous generalization testing trials in this experiment produced poorly differentiated results suggesting that participants found them confusing. Additionally, post hoc examination of the added dimensional values suggested that they actually may not have been perceived as being from the same dimension. As such the continuous generalization trials and the premise diversity and inclusion fallacy effects based on these stimuli were not considered for further analysis in this or subsequent experiments.

Experiment 5

In Experiment 5, the ambiguous testing trials showed a significant preference for responding with the typical non-hidden features over the atypical, ($t(47) = 4.2$, $p < 0.001$). This is consistent with a typicality effect based on non-hidden features. There was no significant preference in responding on the label vs feature trials ($t(47) = 0.5$, $p = 0.607$) or for the label vs hidden feature trials ($t(47) = 0.7$, $p = 0.512$). Overall, this shows no preference for responding consistent with the category label or the non-hidden or hidden features. The generalized classification testing trials showed that 40% of participants responded in one of the four ways that was perfectly consistent with one of the four unidimensional rules (i.e. had errors on the instances that the use of a given rule would have had errors on), matching roughly to the estimate from the error diagrams of 29%.

The classic categorical induction paradigm tests showed a significant effect of conclusion typicality with greater likelihood ratings for the typical conclusion over the atypical conclusion ($t(47) = 2.2$, $p = 0.030$). There was no effect of premise diversity with no significant difference between the ratings for the diverse and less diverse arguments ($t(47) = 0.1$, $p = 0.950$). Likelihood ratings were significantly higher for the general category conclusion than for the specific conclusion ($t(47) = 6.6$, $p < 0.001$), showing the inclusion fallacy. Finally, premise specificity occurred with significantly higher likelihood ratings for the specific premise compared to the general premise ($t(47) = 4.3$, $p < 0.001$).

Experiment 6

In Experiment 6, the ambiguous testing trials showed a significant preference for responding with the atypical non-hidden features over the typical, ($t(47) = 2.6$, $p = 0.012$). This indicates an atypicality effect with non-hidden features. There was no significant preference in responding for the label vs feature trials ($t(47) = 0.5$, $p = 0.627$) or for the label vs hidden feature trials ($t(47) = 1.2$, $p = 0.231$). Again, there was little preference in responding with the

feature typical of the category the label suggested or the feature typical of the category that the non-hidden features or hidden feature suggested. The generalized classification testing trials showed that 23% of participants responded in a way that was perfectly consistent with one of the four unidimensional rules, matching roughly to the reduction in the estimate from the error diagrams to 6%.

In the classic paradigm tests, there was a significant effect of conclusion typicality with the more typical instance being rated as having a higher likelihood than the atypical instance ($t(47) = 3.7$, $p = 0.001$). There was no effect of premise diversity with no significant difference between the likelihood ratings for the diverse and less diverse arguments ($t(47) = 0.05$, $p = 0.962$). There was significantly higher likelihood ratings for the category conclusion than the specific conclusion ($t(47) = 4.9$, $p < 0.001$), demonstrating the inclusion fallacy. Finally, premise specificity occurred as measured by significantly higher likelihood ratings for the specific premise compared to the general premise ($t(47) = 3.9$, $p < 0.001$).

Experiment 7

In Experiment 7, the ambiguous testing trials showed a significant preference in responding with the typical non-hidden features over the atypical, ($t(14) = 2.6$, $p = 0.021$), showing a typicality effect. Participants preferentially responded with the typical feature of the category denoted by the feature information on the label vs feature testing trials ($t(14) = 2.9$, $p = 0.012$). Contrastingly, for the label vs hidden feature testing trials, participants showed a preference for the typical feature of the category denoted by the label information ($t(14) = 3.1$, $p = 0.009$). These results suggest an ordering of preference for using non-hidden feature information above the category labels and a preference for the category labels above the hidden feature information.

In the classic categorical induction paradigm tests, there was not an effect of conclusion typicality as the rated likelihoods of the arguments including the typical and atypical premises

were not significantly different ($t(14) = 0.2$, $p = 0.843$). Similarly, there was no effect of premise diversity with no significant difference between the diverse and less diverse arguments ($t(14) = 0.5$, $p = 0.628$). For the test of the inclusion fallacy, participants gave higher likelihood ratings for the general category conclusion compared to the specific conclusion ($t(14) = 3.2$, $p = 0.006$), demonstrating the inclusion fallacy. Finally, an effect of premise specificity was found with marginally higher likelihood ratings for the specific premise compared to the general premise ($t(14) = 1.8$, $p = 0.094$).

The generalized classification testing trials showed that 20% of participants responded in a way that was perfectly consistent with one of the unidimensional rules. This matches roughly to the estimate from the error diagrams of 38% of participants showing rule like performance during the learning trials.

Experiment 8

For Experiment 8, the ambiguous testing trials showed a significant preference for responding with the typical hidden feature over the atypical, ($t(42) = 3.2$, $p = 0.002$), showing a typicality effect. For the label vs feature testing trials, participants did not consistently, preferentially use the label or the non-hidden features to respond, ($t(42) = 1.3$, $p = 0.212$). For the label vs hidden feature trials, participants showed a prevailing preference for responding with the typical feature of the category denoted by the category label, ($t(42) = 3.1$, $p = 0.004$). This suggests that the hidden features might not have been a preferred basis for responding over the category label.

In the classic categorical induction paradigm tests, there was a conclusion typicality effect in which the rated likelihood of the argument including the typical premise was significantly higher than the likelihood for the atypical premise, ($t(42) = 4.1$, $p < 0.001$). There was no effect of premise diversity with no significant difference between the diverse and less diverse arguments, ($t(42) = 0.1$, $p = 0.959$). For the test of the inclusion fallacy, the likelihood

ratings were significantly higher for the general category conclusion than the specific conclusion, ($t(42) = 5.7$, $p < 0.001$), demonstrating the inclusion fallacy. Finally, there were higher likelihood ratings for the specific premise compared to the general premise, ($t(42) = 5.8$, $p < 0.001$), demonstrating premise specificity.

The generalized classification testing trials showed that 2% of all participants responded on these trials in a way that was perfectly consistent with one of the unidimensional rules. This somewhat matches the estimate from the error diagrams of 23% of participants seemingly using rules on the learning trials.

7.5. Appendix E: All classic paradigm categorical induction questions used in Experiments 5-8, (from Hayes et al., 2010).

Premise Typicality Question 1:

Sparrows have property X / Therefore / Geese have property X

Premise Typicality Question 2:

Penguins have property X / Therefore / Geese have property X

Conclusion Typicality Question 1:

Vultures have property Y / Therefore / Sparrows have property Y

Conclusion Typicality Question 2:

Vultures have property Y / Therefore / Quail have property Y

Premise Diversity Question 1:

Lions have property Z / Mice have property Z / Therefore / Mammals have property Z

Premise Diversity Question 2:

Lions have property Z / Tigers have property Z / Therefore / Mammals have property Z

Inclusion Fallacy Question 1:

Crows have property A / Therefore / Birds have property A

Inclusion Fallacy Question 2:

Crows have property A / Therefore / Ostriches have property A

Premise Specificity Question 1:

Birds have property B / Therefore / Sparrows have property B

Premise Specificity Question 2:

Animals have property B / Therefore / Sparrows have property B