The Incidental Learning of Feature Distributions in Supervised Classification

Jordan T. Thevenow-Harrison

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Educational Psychology)

at the
University of Wisconsin—Madison
2018

Date of final oral examination: 2018-04-20
The dissertation is approved by the following members of the Final Oral Committee:

Charles W. Kalish, Professor, Educational Pscyhology
Martha W. Alibali, Professor, Psychology
Timothy T. Rogers, Professor, Psychology
Haley Vlach, Assistant Professor, Educational Psychology
Martina Rau, Assitant Professor, Educational Psychology

Contents

Acknowledgements v

Introduction 1

Experiment 1 13
- *Methods* 15
- *Results* 29
- *Discussion* 40

Experiment 2 43
- *Methods* 45
- *Results* 47
- *Discussion* 55

Experiment 3 59
- *Methods* 59
- *Results* 61
- *Discussion* 64

Experiment 4 69
- *Methods* 70
- *Results* 72
- *Discussion* 75

Experiment 5 79

Methods 79

Results 81

Discussion 85

General Discussion 87

Table of Results of Experiments 2–5 97

References 101

List of Figures

1 In the inconsistent condition of Experiments 2–5, conflict items are those stimuli that would be classified a different way were the t2 category boundary consistent with the trough of the bimodal stimuli presented during t1 11

2 The distribution of relevant and incidental feature values during training in Experiment 1 17
3 An example of extreme, modal, and marginal values for radial frequency component stimuli on either side of the incidental trough and relevant category boundary 20
4 An illustration of the extremes of the feature space for Gabor patches presented in Experiment 1 21
5 An illustration of the extremes of the feature space for RFCs presented in these experiments 21
6 Accuracy during the training phase of Experiment 1 29
7 RFC Mean reaction time per trial during the training phase 29
8 Experiment 1 RFC accuracy by relevant dimension feature value 29
9 Experiment 1 training accuracy by relevant boundary distance 30
10 Experiment 1 training reaction time by relevant boundary distance 31
11 Experiment 1 RFC accuracy by incidental dimension feature value 31
12 Experiment 1 RFC frequency rating means and confidence intervals 32
13 Experiment 1 Gabor accuracy by relevant dimension feature value 35
14 Accuracy during the Gabor training phase of Experiment 1 35
15 Reaction time during the Gabor training phase of Experiment 1 35
16 Experiment 1 Gabor training accuracy by relevant boundary distance 36
17 Experiment 1 Gabor training reaction time by relevant boundary distance 36
18 Experiment 1 Gabor accuracy by incidental dimension feature value 37
19 Experiment 1 Gabor frequency rating means and confidence intervals 38

20 Experiment 2's feature distributions and category boundaries in the training and testing phases 45
21 Experiment 2 training accuracy by relevant boundary distance 51
22 Accuracy during the t1 phase of Experiment 2 52
23 Reaction time during the t1 phase of Experiment 2 52

24 Accuracy during the t1 phase of Experiment 3 61
25 Reaction time during the t1 phase of Experiment 3 61
26 Experiment 3 training accuracy by relevant boundary distance 61
27 Experiment 3 t2 feature accuracy by feature value of the relevant dimension 62

28 Experiment 4's feature distributions and category boundaries in the t1 and t2 phases 71
29 Experiment 4 t1 accuracy by relevant boundary distance 72
30 Accuracy during the t1 phase of Experiment 4 72
31 Reaction time during the t1 phase of Experiment 4 72

32 Experiment 5's feature distributions and category boundaries in the t1 and t2 phases 80
33 Experiment 5 t1 accuracy by relevant boundary distance 81
34 Accuracy during the t1 phase of Experiment 5 82
35 Reaction time during the t1 phase of Experiment 5 82

Acknowledgements

Thank you, to:

The Department of Educational Psychology at the University of Wisconsin—Madison, which gave me a job teaching people to use their computers, which I now realize *was* a teaching position, and which allowed me to study whatever I wanted, not just what paid.

My mother, Victoria Thevenow, for her unflagging support of my interests, no matter what they are or have been, and her dedication to use what she learned to personally improve the lives of tens of thousands of children.

My advisor, Charles W. Kalish, who gave me a map, showed me the sea, and gave me slightly less rope than would form a noose.

My wife, Dr. Caroline C.-M. Williams-Pierce, for holding a lantern when the sky gets dark, and for illuminating her footprints along the trail.

Linda and Eric Mjoslnes, founders of Bloomington Montessori school, who kindled the flame of the roaring conflagration within me.

Introduction

When we move through the world we constantly learn from examples. What we learn from those examples shapes the categories and representations we use when we make predictions, and we are always making predictions, whether explicitly and intentionally or not. How do those examples shape our predictions? The situations and tasks we engage in shape the categories we use to generalize the knowledge we gained in those tasks to new situations. How do the properties of examples we learn from, their variations, and their distributions come to shape our categories and how we use them? The experiments described in this paper attempt to answer some of these questions.

We present a set of experiments about how the statistical properties of stimuli incidental to a supervised classification task influence later learning. By "incidental" we mean not directly relevant to the immediate supervised classification task (e.g., properties not of the dimension upon which classification is based on the task). After exposure to task-irrelevant but reliably varying features, do people "transfer in" such knowledge to new problems where that information is suddenly applicable, or do they learn nothing at all in the first place such that they have nothing to transfer out of the first learning situation (Broudy, 1977, as cited in Schwartz, Bransford, Sears, & others, 2005)? What information about the irrelevant features bias people when they perform the later task where that

feature becomes relevant? Do they learn about the frequency of feature values (Experiment 1)? Does their exposure to such values interfere with learning a feature distribution, and do they learn to use novel feature values to complete the task (Experiments 2 & 3)? Does the skewness or range of the incidental feature dimensions influence later learning (Experiments 4 & 5)? These experiments all serve to test the hypothesis that in supervised classification with two continuous dimensions the learner does unsupervised learning of the incidental feature.

Sally sorts seashells to make a necklace for her mother. Her father walks alongside her on the shore. They start by picking out good shells for the length of the necklace. Sally picks up a shell and her father either nods or shakes his head; good shells for this part of the necklace should be large, but he says nothing because he is an example character in a psychology paper. Seashells also have noticeable ridges or striations that give texture to their surface: when Sally picks out shells she initially attends to both of these features. Eventually, she recognizes what she should pay attention to is size as it is predictive of her father's approval, while the spacing between the shells' striations is not. Does she still learn anything about the striations? Why would she waste her attention and memory on something with no predictive power? Now imagine the task changes: her father says they should find shells that are good for the centerpiece of the necklace. These are shells with thick striations, but again he says nothing. What, if anything, has discriminating by size prepared her to learn about striation? Is she any different than a seashell-seeking novice when learning about that dimension? All of a sudden the distribution of striations is important, especially so if that feature is bimodally distributed such that the place between the modes, of thickly and narrowly striated shells, could itself form a boundary in her father's approval.

In the task described above, *do* people learn about the spacing of the shells'

striations when it is initially incidental to the task? Rationally, why would they? Because the feature is incidental to the task there is no reason to pay attention to it, to *waste* attention on it. What would be the advantage of using (or diverting) attention to form a representation of the *distribution* of the particular values of the spacing feature? Such distributional information could be useful in some later task; clusters of features might belie some commonality or second-order category that could be useful in the future, so *why not* learn about them? All of these are ways of asking what about task-incidental features, if anything, is transferred out of the traditional supervised classification task and transferred into tasks where those features become relevant. Many models of categorization posit learners' selective attention focuses only on those dimensions useful for categorization (Kruschke, 1992 though see below; e.g., Nosofsky, 1986).

What we seek when we talk about a representation is a description of the relationship between the performer of a task and the task itself. The classic supervised learning paradigm presents the learner with a set of features, usually composed into an object, and requires the learner to respond with the category to which those features correspond. The stimuli are presented one at a time, and after each presentation and response, the learner receives feedback. Experimenters assess learning above and beyond mere memorization by presenting the learner with examples in a second phase (t2) that were not present in training (t1) but fit the learned rule or boundary. Work done in the past operated under the assumption that this task fostered representations that applied to a variety of natural contexts (Markman & Ross, 2003). But the task's sequential feedback after each guess encourages the learner to focus on providing guesses to test hypotheses that only serve to discriminate between the categories, with little to no emphasis on information about the composition of the categories themselves. The traditional cat-

egory learning task is probably not a good representation of the kind of learning that happens in the real world, where categories rarely possess such properties, and the modes of learning we engage in vary. One criticism of the supervised classification task is that it amounts to training a discriminative classifier, which requires only the information to make the classification and does nothing with any other information (i.e., it only learns what is necessary to make the discrimination or classification). In a paradigm almost exactly like the classic supervised classification task but where participants only observe the stimuli then receive feedback rather than making the classifications themselves, Levering & Kurtz (2014) found that participants learn more about the distributional properties of the features and categories that make up the task. Other research into new learning tasks has emerged from this criticism in the last 30 years (Love, 2002; Markman & Ross, 2003; Wattenmaker, 1991; Yamauchi & Markman, 1998). But this dissertation seeks to probe what people learn about the incidental feature in the supervised classification task.

Categorization allows us to organize and extend our knowledge to make predictions about new things in the world. Some features of categories are more useful for these kinds of predictions than others. If we see a ball bounce high and we are asked to predict whether a new ball will do the same, the material the ball is made of may be a more useful feature to attend to than the ball's color. Most theories of learning involve this kind of selective learning component (Kruschke, 1992; Love, Medin, & Gureckis, 2004; Medin & Shaffer, 1978; Nosofsky, 1986).

Levering (2012) describes a framework that draws heavily from distinctions within the machine learning literature. Classification models of two types are used extensively in machine learning: discriminative and generative (e.g., Ng & Jordan, 2001). The two are distinguished by whether or not they include more information than they need to discriminate between categories in a learning task.

Discriminative models do not preserve such information, while generative models do. Discriminative models focus solely on the task of classification; they do not represent the properties of their categories' or features' distribution, such as range or variance. Generative models do account for learning such properties as they are required to generate predictions as to what category was most likely to yield a given input. They solve a more general problem than the task of only classifying input. If we apply the ideas mentioned above to human performance on category learning tasks, purely discriminative category learning should yield very accurate and fast category learning but would lack information outside of the categories themselves. Purely generative category learning would yield representations of categories that fully preserve the features (and their distributions) present in the training sample, not only of features diagnostic to the classification task but also non-diagnostic features as well.

A popular theory in category learning claims we learn by attending only to those dimensions that are most diagnostic (Kruschke, 1992, 2003; Kruschke & Johansen, 1999; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). A categorization problem is easier to learn the fewer diagnostic dimensions it has (Shepard et al., 1961) and when the task emphasizes those diagnostic dimensions through blocking and highlighting (Kruschke, 2003; Kruschke & Blair, 2000). After learning a particular dimension is not diagnostic, it is hard to switch to using it as diagnostic (Kruschke, 1996). The degree to which a feature dimension is diagnostic of category membership is correlated with the amount of time people spend looking at it (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005).

Selective attention helps learners focus on category-relevant dimensions, especially in situations where stimuli have few features in common. When the task is to generalize, selective attention is particularly useful for focusing on relevant

dimensions and ignoring irrelevant ones (Best & Yim, 2013). Selective attention also comes with costs: learned inattention (Kruschke, 1992) is when a learner continues to ignore information that was previously irrelevant—that is, they learn to ignore. This can be a cost when a new learning situation involves previously irrelevant information, and illustrates a distinction between the cost of switching tasks, where a rule or category boundary might have changed but was hard to notice, and the cost of not learning about a dimension that was previously incidental to the task but is now relevant. The cost of selective attention goes down when a new category to be learned uses the same dimension as the previous one but with different values (Best & Yim, 2013). This is called an intra-dimensional shift. An extra-dimensional shift is when the category-relevant dimension(s) change. Adults show learned inattention during extra-dimensional shifts between categories in category learning tasks (Dopson, Esber, & Pearce, 2010; Kruschke & Blair, 2000, Hoffman & Rehder, 2010). That is, they attend selectively while learning the first category and incur the cost of inattention to the previously irrelevant dimension while learning the second category if an extra-dimensional shift was present (Best & Yim, 2013). However, extra-dimensional shifts can be easier than reversal shifts—where the classifications switch but the relevant dimension stays the same—if the newly relevant dimension contains novel feature values not seen before the shift (Kruschke, 1996).

These findings seem to situate supervised classification as purely discriminative. But after making same-different discriminations between faces morphed along two dimensions, undergraduates were sensitive to task-irrelevant distributional structure within categories (Gureckis & Goldstone, 2008). Many of the stimuli in the above experiments are spatially separated such that their features are structurally independent from one another. The learning "impairments" seen in

supervised classification learners when compared to other kinds of learning tasks may not manifest when stimuli are integrated or rich (Hoffman & Rehder, 2010). Participants learned about environmental information incidental to the learning task in a rich scene where learners searched for animals with varying features in different places (an effect that went away when the position of the animals was fixed, Allen & Brooks, 1991). Participants in supervised classification tasks learn more about stimuli when they are predicting a category label than a supposedly meaningless outcome (e.g., a high or low tone, Bott, Hoffman, & Murphy, 2007). They are more sensitive towards nondiagnostic atypical features than in other learning paradigms (Jee & Wiley, 2014). Therefore it is not clear what or when people learn about non-diagnostic information in the supervised classification task. Perhaps the supervised classification task is more generative than previously thought.

In her dissertation, Levering (2012) compares observation with the traditional supervised classification task. Her second experiment presented learners stimuli that vary continuously (e.g., 12 values instead of the binary dimensions in most of the literature on supervised classification) along two dimensions, one diagnostic of category membership and one dimension that was either partially diagnostic or non-diagnostic dimension. After either a supervised classification training phase or a phase very similar to it but strictly observational, participants were asked to make typicality judgments for given stimuli. Rather than select the most typical of two examples, participants rated each presented stimuli and category label based on its typicality of that category. Levering then subtracted the average typicality rating of stimuli not present in training (with more extreme feature values) from the average typicality rating of examples shown in training (with a "central" feature value). Along the diagnostic dimension, observational learners had higher difference scores than classification learners, though both were above chance.

However, the partially diagnostic dimension showed an interesting finding: Both observational and supervised classification learners had typicality difference scores above chance, though the observational learners had higher typicality differences scores than classification learners, indicating that both conditions learned something about the frequency of the partially diagnostic dimension, even when presented with a dimension that was completely diagnostic. In the condition where the dimension was entirely non-diagnostic learners in both conditions had typicality difference scores above chance and not significantly different from one another. But when they performed a kind of feature inference task where they were given the category and asked to predict the feature, neither group showed an awareness of the distribution of features along the non-diagnostic dimension. In summary, both classification and observational learners do well at supervised classification, and while they do learn *something* about the non-diagnostic dimension (inasmuch as they are above chance in typicality judgments, but not good at them), they do not have a sense of the distribution of non-diagnostic features, of what constitutes a typical or atypical feature.

Overview of Experiments

This study focuses on what people learn about feature distributions in supervised classification. Most research in categorization has focused on binary dimensions (Shepard et al., 1961). Previous research found that people learn about non-diagnostic but characteristic dimensions (e.g., Allen & Brooks, 1991; Brooks, Squire-Graydon, & Wood, 2007), but that literature focuses on binary or non-continuous stimuli and has not focused on how that information is used later. Experiment 1 is a conceptual replication of this kind of research with stimuli that have more feature values and are distributed in a particular and intentional way. Research in the realm of reversal, intra-dimensional, and extra-dimensional shifts

in categorization examined how hard it is to use previously task-incidental features, but has not focused on continuous dimensions, or how the distributional structure of the previously irrelevant dimension influences judgments after the shift occurs (e.g., Kruschke, 1996). Many of the aforementioned studies in the traditional supervised classification task use stimuli with perceptually separable dimensions (e.g., rotation and spacing in Gabor patches) rather than perceptually integrated dimensions (e.g., saturation and brightness), particularly those that suggest people ignore incidental dimensions. As a part of our search for evidence of learning about task-incidental dimensions, we think we are more likely to find this learning with integrated stimuli (Hanania & Smith, 2010 for reviews on these distinctions and their development; see Smith & Kemler, 1978; Ward, 1980).

Do people learn about the distribution of the incidental feature in a categorization task? The first experiment presented a training phase where participants categorized a series of stimuli with two uncorrelated features where the classification rule was unidimensional (e.g., a division between feature values along a single dimension). This experiment examined whether people are aware of the distribution of the features during a supervised classification experiment, and in particular whether they are aware of the incidental features. This also served as a conceptual replication of Chin-Parker & Ross (2004) and Levering (2012).

If we do find evidence of learning about the incidental feature dimension, just *what* do people learn? Another thing besides typicality that participants might learn about the incidental feature in a supervised classification task is a general sense of the distribution. Perhaps they do not learn about frequency in the above experiment, but instead learn about distributional features like troughs, where they saw no examples of a feature value, or simply are able to rely on novelty to drive category inference (e.g., have they seen a given stimulus before). After a

training session, Experiment 1 tried to measure what feature values for stimuli participants thought were more or less frequent. It also presented a forced-choice task to compare the perceived frequency of one dimension's feature values while holding the other constant.

The second experiment asks whether people use what they learn about feature distributions during this first task described above in a second task. Participants completed a first task (hereafter called t1) similar to the one described in Experiment 1. Then, in a second task (hereafter called t2), participants performed a task much like the t1 task, but where the previously incidental feature became the relevant feature and vice versa. In t2 the feature distributions changed from a bimodal to a uniform distribution. Participants were assigned to a *consistent* or *inconsistent* condition where the consistent condition had a category boundary in the same place as the trough in the incidental feature distribution present during t1 when the now-relevant feature was incidental to the task. In the inconsistent condition, the location of the t2 category boundary was at one of the modes of the bimodal distribution of the previously incidental feature distribution. If participants in the inconsistent condition performed worse on the categorization task than those in the consistent condition then there is some evidence of cross-task transfer of incidental information. We also present and test some alternative hypotheses about what could be learned in such a situation. Experiment 3 refined Experiment 2 by explicating the number of categories and point in the experiment that the relevant dimension changes from one feature to the other. Experiment 4 tried to minimize the differences between t1 and t2 stimuli by making t2 stimuli basically the same (both bimodal) at t1 while only changing the category boundary. Experiment 5 was the culmination of this experimental design wherein the consistency condition was decoupled from the actual feature value along a stimuli's dimension and

tied only to the consistency between t1 and t2.

At the root of this dissertation is the hypothesis that in supervised classification of one dimension when stimuli have two dimensions the learner does unsupervised learning of the incidental feature during t1. In the case of Experiments 2 through 5, there are two clusters with a trough in the middle—little ones and big ones. Now the learner can represent that as a cluster structure. Perhaps the most interesting comparison happens within the inconsistent condition of Experiments 2–5, the condition wherein the category boundary shifts during t2 from where a trough was presented during t1 to the feature value where one of the modes was during t1. A hypothesis of these experiments is that in the t2 phase, boundaries that respect this cluster structure will be easier to learn than boundaries that violate it. What it means to violate that cluster structure, to be misaligned with it, is the case when stimuli with relevant feature values in the same cluster end up on different sides of the category boundary. Of course, there is no inconsistency in this cluster structure for participants in the consistent condition: their cluster structure maps onto the category boundaries present in the t2 phase. These "conflict items" in the inconsistent condition will be an important diagnostic tool.

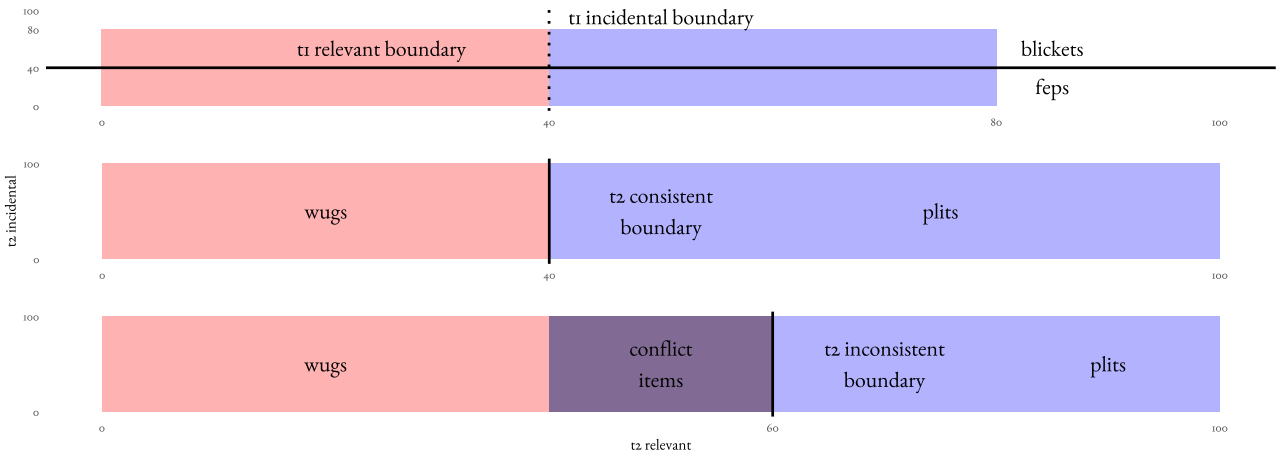

Figure 1: In the inconsistent condition of Experiments 2–5, conflict items are those stimuli that would be classified a different way were the t2 category boundary consistent with the trough of the bimodal stimuli presented during t1.

What is generalized from one categorization task to another? It is possible

that more general properties of the feature distributions during the first task might drive performance during the second task. If the categories in the first task are of equal size perhaps that information is transferred to the second task and informs participants' classifications there. Even if participants do not learn about the frequencies of the feature distributions, perhaps they learn something more general about the range of the feature clusters themselves. The fourth and fifth experiments will manipulate the range of the incidental feature distribution during t1. How will participants who are exposed to such a distribution do when the incidental feature becomes relevant in a new classification task, and the new relevant distribution is either consistent or inconsistent with what was seen during t1? Experiments 2–5 are akin to studies focusing on extra-dimensional shifts, exploring whether these shifts are hard for learners in supervised classification, and looking at what information, if anything, do learners bring over about the feature distributions from the first classification task to the second.

Experiment 1

Experiment 1 examines how people in a supervised classification task learn about the distribution of feature incidental to the task. Participants will classify stimuli composed of two continuous bimodal dimensions where one dimension is relevant to the task, the other incidental and uncorrelated with the relevant dimension. Then they choose the most common of two stimuli with values that vary in one dimension, either the incidental or relevant feature dimension. This experiment also serves to help us design and validate stimuli within a similar experimental paradigm as Chin-Parker & Ross (2004) and Levering (2012).

As a part of their first experiment, Chin-Parker & Ross (2004) presented undergraduates stimuli with five binary features and asked them to classify them into one of two categories. Then they were asked to choose the "most typical" of two stimuli that matched either in prototypicality (e.g, number of features matching an unseen category prototype established during 11) but varied in diagnosticity (e.g, number of features that are actually diagnostic of the category), or vice versa. They also asked participants about their confidence in their judgments, which the experimenters multiplied with accuracy into confidence-adjusted scores. Classification learners based their judgments of typicality on the diagnosticity of the stimuli alone; the number of features shared with the category prototype did not seem to influence their choice of the most typical member of the category. Chin-

Parker and Ross interpreted this as evidence that supervised classification learners either did not learn about non-diagnostic (but prototypical) features during t1 or did not use that information when making judgments of typicality. Another condition in the same experiment presented learners with stimuli that included the category label but were missing a single feature. These "inference learners" did not focus exclusively on diagnosticity the way supervised classification learners did. Inasmuch as they are similar, the "incidental" dimension in this dissertation aligns somewhat with the aforementioned non-diagnostic but prototypical dimension. In our experiment, our non-diagnostic dimension does not have this prototypical structure.

Both Chin-Parker & Ross (2004) and Levering (2012) examined how sensitive supervised classification learners were to learning about feature dimensions that were incidental to a unidimensional category boundary. Both found that supervised classification learners do not show much evidence of learning about this dimension when it is entirely non-diagnostic. However, the methods the researchers use to determine this use simpler dimensional representations than this dissertation uses. Chin-Parker and Ross use binary feature dimensions and make their conclusions based on the *number* of features, rather than those features' frequencies along a distribution. Levering's typicality difference scores compared trained and untrained feature values alone rather than manipulating the frequency of those values such that certain feature values are more common than others. In this study, we use two sets of stimuli between participants, one set with perceptually separable features, and one with perceptually integrated features. We believe the perceptually integrated features will afford more of an opportunity for participants to learn about the incidental feature distribution, but for this experiment, we contrast the results between the stimuli sets as a kind of pilot for the later ex-

periments. The stimuli are discussed in more detail in the following section. The first experiment in this dissertation examines whether or not frequency influences learning about non-diagnostic features.

Methods

The studies described in this proposal was conducted on Amazon Mechanical Turk. Crowdsourcing is the process of using a large number of people to complete a task in parallel, usually online. In the last few years, it has grown in popularity with social scientists seeking to increase the number and diversity of participants in their experiments. Crowdsourcing platforms allow studies that would have taken weeks or months to be performed overnight (Litman, Robinson, & Rosenzweig, 2015).

Amazon Mechanical Turk (MTurk) is one of if not the most popular choice for crowdsourcing social science experiments. Within psychology, MTurk has been used to study clinical interventions, political theories, body image, game theory, religious beliefs, opinions on personal philosophy, and education, among many other fields. Currently over 500,000 people from all over the world routinely complete tasks on MTurk. Their backgrounds are more diverse in terms of age, education level, ethnicity, income, and gender than typical undergraduate samples (Berinsky, Huber, & Lenz, 2012).

Crowdsourced data has a few factors that influence the validity of data. Unlike a laboratory setting, it is not possible to tightly control the environment in which data is collected. Chandler, Mueller, and Paolacci (2013, as cited in Litman et al., 2015) showed that among participants with approval ratings over 95%, 18% of participants on MTurk watch TV and 6% use instant messaging while completing tasks. As odd as it may intuitively seem, monetary compensation has little to no

influence on data quality. The only influence monetary compensation seems to have is on the speed of recruitment, where lower rewards result in slower data collection, but not data quality (Buhrmester, Kwang, & Gosling, 2011). Why? MTurk workers were asked to rank the importance of the following motivations: killing time, making money, having fun, participating in interesting tasks, and gaining knowledge. They reported that task enjoyment is the most important motivation, followed by killing time and having fun. Making money only rated above gaining knowledge (Buhrmester et al., 2011). This finding seems to indicate that MTurk workers are intrinsically motivated. However, more recent studies have shown that monetary compensation has become the primary motivator to workers, even those based in the US, though this change in motivation has not affected data quality. US workers are less interested in low-paying tasks, so studies that target them that also want fast turnaround tend to pay more than the average MTurk hourly rate (Litman et al., 2015).

MTurk is a dynamically changing workforce influenced by quickly changing market forces. Near its inception, workers were predominantly from the US; now only 50% of workers are located in the US, and 40% are from India. The differences in motivations between these two dominant demographic groups are highlighted by their motivations: in India, completing tasks on MTurk are workers' primary source of income. Overall, MTurk workers are young, between 21 and 35; mostly female, about 70%, and have lower household incomes, with 60% earning less than $60,000 annually. Most workers complete between 20 and 50 tasks per week (Ipeirotis, 2010).

The experiments presented in this proposal were only made available to MTurk workers located in the United States with approval ratings over 95%. The framework for the experiment structure is designed with the jsPsych library, which

simplifies the creation of web applications for online psychology experiments (Leeuw, 2014). The integration between the experiment and MTurk was managed by PsiTurk

[1]

[1] https://psiturk.org

which handles data exchange, structures the flow of the experiment between MTurk's servers and the experiment server, and coordinates participants such that no one participates in the experiment more than once and participants' anonymous identifiers are matched and checked with their data.

Participants. Eighty-one participants completed the experiment on Amazon Mechanical Turk. Of those, three were removed for reloading the experiment. 78 participants are included in this analysis. 39 participants were trained on radial frequency component stimuli, and 39 on Gabor patch stimuli.

Design. Experiment 1 sought to understand whether and how people learn about the distribution of a feature dimension incidental to the category structure present in a supervised classification task. The experiment consisted of three tasks: the supervised classification training phase, the forced-choice task, and the frequency rating task. The training phase was a supervised classification task wherein participants saw a two-dimensional stimulus then classified it into one of two categories. The forced-choice and frequency rating tasks tested what the participants learned during the training phase. The training phase was presented first. The forced-choice task and the frequency rating task were presented in an order that varied between participants.

First, we trained participants on two-dimensional stimuli. The training phase was a standard supervised classification task that presented participants with 80 stimuli that varied bimodally across two dimensions. The instructions asked

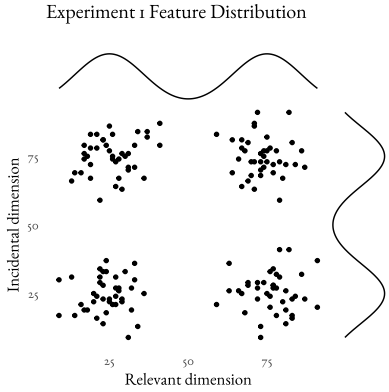

Figure 2: The distribution of relevant and incidental feature values during training in Experiment 1. Each point represents a stimulus. Each dimension is bimodal and uncorrelated with the other.

them to press a button corresponding to one of two named categories after seeing the stimulus. After each response, the participant received feedback indicating whether their answer was right or wrong, and if their selection was wrong they saw the correct category label. Category membership was based on one dimension, the relevant dimension, which was counterbalanced between participants. The distributions of the relevant and incidental dimensions were the same during training but uncorrelated to prevent any incidental feature range from associating with a category label. Both dimensions of the stimuli were distributed bimodally. No stimuli had feature values between the tails of these modes. One category consisted of items with a relevant feature value up to and including 50, and the other category consisted of stimuli with a relevant feature value between 51 and 100, in order to present an equal number of category members during the task. (See the Materials section for more detail on what these numbers mean.) Participants classified 80 stimuli.

Next, we sought to evaluate what participants learned about the distribution of the relevant and incidental dimensions through a frequency rating task and a forced-choice task. These serve as a way of gauging whether or not participants learned anything about the *distributions* of the relevant and incidental feature dimensions, above and beyond what is necessary (e.g., the location on the relevant dimension of the category boundary). The frequency rating task presented participants with a subset of the stimuli they saw during the training phase. Participants were first briefed about what "typical" meant in the context of this experiment and were given examples contrasting something that is highly representative of a category (a robin) against something that is not representative of the same category (a penguin). During each trial, they saw a stimulus with dimensions composed of either extreme values (the outer tails of the distribution), marginal values

(the inner tails of the bimodal distributions nearest the trough), or modal values. There are two possible values for each of these, so each dimension could take on one of six possible values. Participants saw every combination of these feature values once for a total of 36 trials. Participants were then asked to rate how "typical" a stimulus was on a Likert scale from 1 to 7, where 1 was labeled "not at all typical", 4 was labeled "moderately typical," and 7 "highly typical."

The forced-choice task also presented participants with a subset of the stimuli they saw during the training phase. Participants saw two stimuli, one on the left and one on the right, and were asked to pick the one they saw more of. Participants made 16 choices in this phase. For both stimuli, one dimension was fixed at a random value shared by the two stimuli, while the other dimension of the two stimuli held two different values described in the previous task: extreme, marginal, or modal. One stimulus had a modal feature value, while the other had a feature value in one of the extreme or marginal tails. This made it so there was always a right answer for every question, rather than present participants with questions forcing a choice between two tail values, for which there are no right answers.

Materials. This experiment will use Gabor patches and Radial Frequency Components (RFCs). Gabor patches are sinusoidally striped patches of light and dark gray centered in Gaussian noise. Gabor patches are useful here because their features are perceptually separable, in that their rotation (orientation of the light and dark lines created by the sinusoidal interference pattern) and spacing (the distance between those lines) features are perceived immediately and distinctly. However, as a single visual stimulus one feature could not exist without the other. The stimuli are 400 x 400 pixels in size. The x and y values map to coordinates of pixels in the image. The formula outputs real values normalized around 0. These are scaled up to become brightness values between 0 and 255.

Radial Frequency Components are a series of sine waves of different frequencies that describe distortions of a circle (Zahn & Roskies, 1972, as cited in Drucker & Aguirre, 2009). In these experiments they are generated using the same methods described in Op de Beeck, Wagemans, & Vogels (2001) and Drucker & Aguirre (2009).

Source code for generating RFCs was adapted from Daniel Drucker's Ph.D. thesis available at https://github.com/dmd/thesis. Their generation functions are considerably longer than those for Gabor patches.

At one point thought to be the basis for shape recognition (Schwartz et al, 1983, and rejected by Albright & Gross, 1990, as cited in Drucker & Aguirre, 2009), these shapes can be thought of as existing in a two-dimensional space of shape features and the orientation of features within the shape.

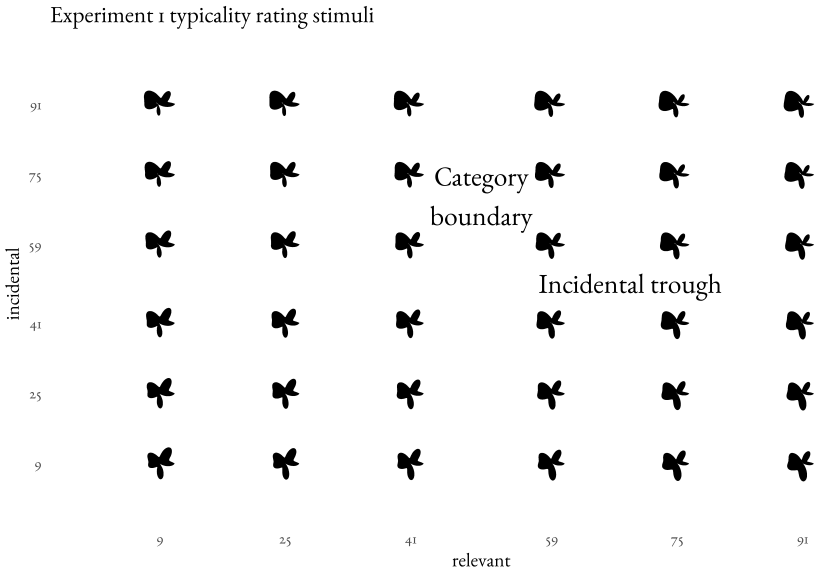

Figure 3: An example of extreme, modal, and marginal values for radial frequency component stimuli on either side of the incidental trough and relevant category boundary.

RFCs are useful because the shapes don't have the kinds of "semantic boundaries" some stimuli like Gabor patches have (Drucker & Aguirre, 2009), where horizontal or vertical orientations of the lines formed by the differentially shaded regions act as reference points from which a participant could judge the angle of rotation, or could count the number of light or dark patches to judge frequency. The features in RFCs are integrated in the sense that one cannot examine one dimension without perceiving the value of the other, a property that we think

should increase the chances of participants learning about the incidental feature distribution if they learn anything about the incidental dimension at all.

Furthermore, in the RFC stimuli, there are no feature properties to count or measure with reference to any baseline like in the Gabor example above (e.g., with RFCs there is no equivalent to a reference point such as a horizontal line in Gabor patches). In this study, we talk about feature values ranging from 1 to 100. These numbers represent a scaling of the actual parameters used to generate the stimuli to make it easier to talk abstractly of feature values regardless of the feature in question. Instead of describing a stimulus with a rotational value of the dark lines in the gradient of 45 degrees and a spacing parameter of .05, we would say this stimulus has a rotation of 50 and a spacing of 50.

It should be noted that there are some issues with this approach as it assumes congruity between psychological or subjective feature space and the feature space used to generate the stimuli. People do seem to treat line orientations of 0 and 90 degrees differently than intermediate values (Xu, Zhu, & Rogers, 2012). In the experiments described in this study, we rarely expose participants to these boundaries in the feature space and opt instead to present oblique angles for the most part.

When we talk about bimodal distributions we mean features that have been generated by sampling a set of numbers from a normal distribution then adding those numbers to the "maximum" number of the range. In the case of Experiment 1's training phase, the feature values range from 1 to 50 with a mean of 25 and a standard deviation of 7. We take that set of numbers then add 50 to them to make the other curve of the symmetric bimodal feature distribution. This provides a natural *trough* in the case of the incidental feature, or a *category boundary* in the case of the relevant dimension, to delineate category membership (or potentially

These Gabor patches are described by $g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi \frac{x'}{\lambda} + \psi\right)$, where λ ranges from .01 to .1, θ ranges from 0 to 90, $\psi = 0, \sigma = 50, \gamma = 1$. Degrees of rotation corresponds with the value of θ and the spacing between the stripes corresponds with λ.

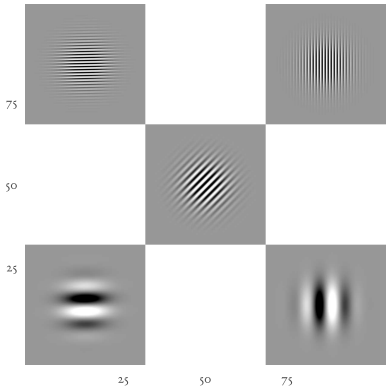

Figure 4: An illustration of the extremes of the feature space for Gabor patches presented in Experiment 1.

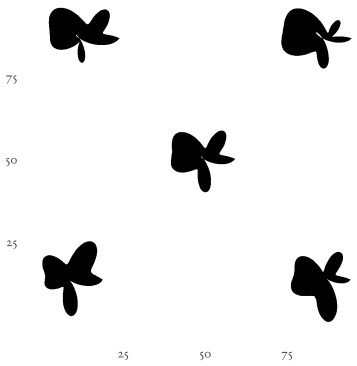

Figure 5: An illustration of the extremes of the feature space for RFCs presented in these experiments.

obstruct it in the case of the t2 phase of the inconsistent condition in Experiment 2). When we say the range of a bimodal distribution is 1 to 100, due to the low probability with which the tails are sampled, the minimum and maximum values are 8 and 92, respectively. All participants saw the same distribution of feature values.

Procedure. After accepting the experiment's ad on Amazon Mechanical Turk, participants are directed to the experiment server. There they consent to participate in the research and are given the following instructions:

> In this experiment you are going to see some pictures. Your task is to determine whether or not each of those pictures is a Fep. You'll learn what a fep is based on the feedback you get answering questions. It's okay to guess for the first few. The pictures appear one at a time. If you think the picture is a fep, press the letter F on the keyboard as fast as you can. If you think the picture is a blicket, press the letter K key on the keyboard as fast as you can. If you think it is a fep press the F key. If you think it is a blicket press the K key. At first you will just be guessing, but after a few you should start to have an idea of what feps and blickets are. The experiment goes faster when you get the questions correct. There are 80 questions in this part. You will have 10 seconds to respond to each question.

In the training phase, participants see an image centered in their browser window. Centered below the image is the prompt "Is this a fep or a blicket?" Below that are instructions for responding: "Press F if this is a fep. Press K if it is a blicket." A judgment is marked incorrect if the prompt is not answered within 10 seconds, long enough that it does not put a premium on perceptual filtering but short enough that a participant cannot abandon the task. Participants categorize 80 unique stimuli this way, then move to the t2 phase, which consists of a forced-choice task and a frequency rating task in an order counterbalanced between participants.

In the forced-choice task participants see two stimuli and are asked to select the one that was more typical. The stimuli are centered on the left and right half of the browser window. Below the stimuli will be the prompt, "Pick the one you saw more of." Participants make 16 judgments this way. They are briefed on the following text:

> Now you're going to do a slightly different task. You will see two pictures similar to the ones you saw before. Your task is to pick the one you saw more of. If you think you saw the picture on the left more often, press the 'Q' key on the keyboard as fast as you can (it's in the upper left of the keyboard). If you think you saw the picture on the right more often, press the 'P' key on the keyboard as fast as you can (it's in the upper right of the keyboard). You won't receive feedback about your answers in this part. There are 16 questions in this part. You will have 10 seconds to respond to each question.

In the frequency rating task, participants make 36 frequency ratings of individual stimuli on a Likert scale from 1 to 7. Participants are presented with a single stimulus above a vertically oriented Likert scale. They are first briefed on the following text:

> Now you're going to do a slightly different task. You will see a picture similar to the ones you saw before. You saw a lot of feps and blickets. There was a lot of variation among the different feps and blickets; some were very common, and some were rare. When you see the pictures in the next task, think of the most common fep and blicket when considering your response. Your task is pick how typical that picture is for its category using the choices below the picture. By typical we mean how common or representative. A robin would be a very typical bird because most birds are like robins. A penguin would not be a typical bird because there aren't many birds like penguins. You won't receive feedback about your answers in this part. There are 36 questions in this part.

Predicted Results. The purpose of this experiment is to see whether

or not people are able to first learn two opposing categories in different sets of two-dimensional stimuli, one perceptually integrated (RFCs), and one perceptually separable (Gabor patches), then use what they learned to answer questions about the distribution of feature values they saw during training. The primary hypothesis is that people will learn something about the incidental feature due to its intentionally designed distribution. Will people learn anything at all about the incidental distribution, and if they do, will they notice the trough in the incidental feature distribution?

Training. Evidence of learning of the relevant distribution during training will be a positive change in accuracy with each training trial. We use trial accuracy as a proxy for learning over time. Pilot data indicated that this slope will be positive and the average participant accuracy will be at least 80%.

How do people infer which category a given stimulus belongs to? We will attempt to answer this question by performing a generalized linear mixed regression using accuracy as the dependent variable and the values of the relevant and incidental dimension as independent variables, along with the trial number to tease out effects of practice or exposure. In the case of Gabor patches, we will add a factor for the relevant dimension (rotation and spacing) because pilot testing indicated that accuracy on those two dimensions can differ. We will include this analysis for RFCs, though previous research indicates their dimensions are of equal salience (e.g., Drucker & Aguirre, 2009).

If participants learn the relevant dimension well enough (or at least learn the category boundary) to accurately categorize stimuli, a model consisting of the relevant dimension (and a random slope for each participant) should fit the data better than a model consisting of only the incidental dimension. A model consisting of both relevant and incidental dimensional values should not be more predictive

of accuracy than the model consisting only of the relevant dimension, because the relevant dimension is, by definition, the dimension that determines a stimulus's category. If the relevant-only model best fits the data we can infer that participants were indeed using the relevant feature values to make predictions about category membership and learned something about the relevant dimensions' distribution, at least enough to determine the boundary for category membership. This analysis will act as confirmation that people are indeed learning and using the relevant dimension, indicating that at least insofar as replication of the classic supervised classification study is concerned, we will have succeeded.

To see how well participants are able to classify stimuli based on the feature values of those stimuli, we will examine the relationship between classification accuracy and feature value. Because the distributions of feature values are binomial along both dimensions, with potentially lower accuracy near the category boundary, instead of using feature values themselves we will use the distance of the relevant feature value from the category boundary for the relevant independent variable, and incidental feature value's distance from the trough of the incidental dimension's distribution, called the boundary distance. This transformation should give accuracy and boundary distance a somewhat linear relationship appropriate for analysis using the methods described above, rather than the nonlinear shape that could come from using feature value directly.

Frequency ratings. Participants rated the "typicality" (the proper term in the literature would frequency) of different stimuli on a Likert scale from 1 ("Not at all typical") to 7 ("Highly typical"). The stimuli consisted of relevant and incidental features at either of the modes of the bimodal feature distributions, extremes (the limits of the feature space) or margins (closest to the trough or category boundary). Participants rated all combinations of these values for a total of 36 ratings

per participant. Because we cannot hold to the assumption of ANOVA or even GLMM wherein the distance between the ratings of the Likert scale are assumed to be equidistant from one another, we perform a Bayesian ordinal logistic regression using the `brms` package in R (Bürkner, 2017).

If participants learned about the distribution of feature values along the relevant and incidental dimensions, modal values should be garner higher ratings than extreme or marginal values; feature values are more frequently distributed around the modes. But there are other things a participant in a supervised classification task could learn; the only real requirement for success in supervised classification is knowledge of the category boundary; any learning about the distribution of even the relevant dimension is not necessary, and absolutely nothing is needed from the incidental feature distribution to successfully complete the task. In the case of the RFC stimuli, however, it's not always clear which dimension is which; some learning about the distributions of both feature dimensions would be expected in participants trained on RFC stimuli. If frequency ratings are unrelated to incidental feature values, this would be evidence that people either did not learn the incidental feature values or did not subsequently associate them with the frequency of a given stimulus.

An alternative hypothesis is that participants form a prototypical representation of the incidental feature dimension: they form a kind of unimodal or single-cluster representation of the distribution of the incidental feature; that is, without a sense of that feature's distribution. They learn enough to distinguish between extremes; extreme values have definitive category membership because of their distance from the boundary. But marginal values may not be represented or remembered as well as the other two value types (or at all) due to their proximity to the category boundary; smaller distances between members of each category

make distinguishing between the membership of marginal values more difficult than other comparisons.

Alternatively, marginal values could be the most well-represented values: because of the difficulty in determining category membership, marginal values could garner the most attention during learning. A confounding hypothesis parallel to this one is that it is possible that participants learn the range, then represent values near the middle of the range as the most frequent because they assume a normal distribution, or their representation of the feature dimension is averaged over category membership such that they learn a prototype model of the dimension, and not a model for the categories themselves. Because the frequency rating task asks only about the frequency of marginal values (41 and 59), not the unseen central values (participants saw no feature values between those marginal tail values of the bimodal feature distribution) it is not possible to differentiate between these hypotheses in this experiment.

Forced-choice frequency comparisons. Participants also completed a forced-choice frequency comparison task, where they were presented two stimuli that shared one feature value and differed in another. The ultimate purpose of this evaluation is to confirm that people will be more accurate on questions where the relevant dimension varies than on questions where the incidental dimension varies. The differing feature values compared marginal and extreme values to modal values, so there was always a correct answer. (Both extreme and marginal values have the same frequency as the entire feature distribution consists of two symmetric normal distributions around the modes.) As such these questions have correct answers: always choose the modes. However, if the marginal values are the most salient and are considered to be most frequent because of the hypothesis described above, it may be that participants selected marginal values more often

than modal ones.

To examine accuracy we will perform a generalized linear mixed model with accuracy as the dependent variable and the alternative feature value that varies between the two presented stimuli as an independent variable (modal values are presented opposite marginal or extreme values) and an independent variable for dimensional relevance (whether the varying dimension is relevant or incidental), with a random slope for each participant. If participants are learning the distribution of features of both relevant and incidental dimensions equally then dimensional relevance should not influence model fit. However, the literature suggests that participants will have learned more about the relevant dimension as it determined category membership during the training task, so we expect that participants will be more accurate on relevant than incidental comparisons. Indeed, if they learn nothing about the incidental dimension during training then the factor of dimensional relevance should contribute to the model fit, such that when the dimension that varies is relevant accuracy is higher. Furthermore, if participants' representations or memories of the dimensions include information about frequency then there should be no difference between accuracy for different values of the alternative feature. If, however, they are representing these features differently, perhaps using a prototype representation where marginal (more central) values are considered more frequent, we suspect the factor for alternative values will affect model fit. An interesting comparison will be the simple effects within dimensional relevance for the alternative feature values: if the existing literature is correct that people only learn what is necessary to complete the supervised classification task then there should be no interaction. If, however, accuracy for the comparisons of alternative feature values are equal in the relevant dimension and unequal in the incidental dimension then that is evidence of some kind of learning for the

distribution, or at least the range, of the incidental dimension.

Results

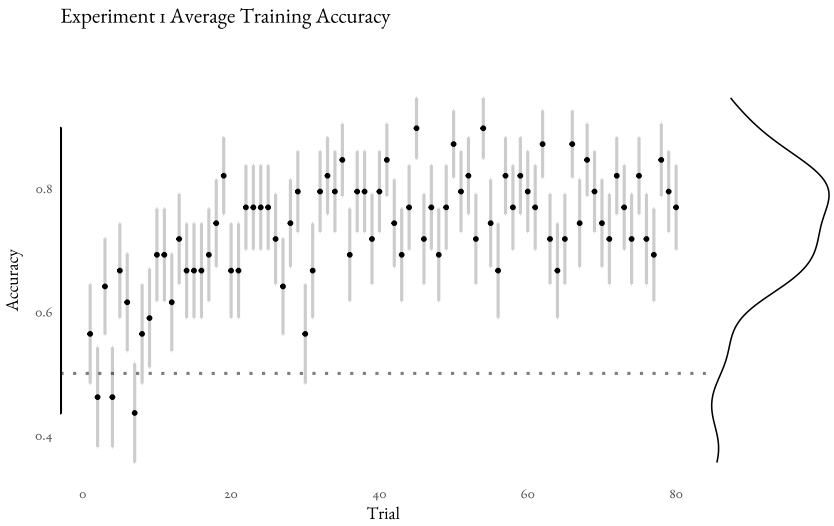

Figure 6: Accuracy during the training phase of Experiment 1. Points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance. Accuracy improved over training but never reached ceiling.

RFC Training. Participants were accurate during training. Calculating an average accuracy score for each participant, we see $\mathcal{M}_{\text{m}} = 0.73$, $SE_{\text{m}} = 0.05$. Overall mean accuracy was above chance in a one sample t-test, $t(38) = 12.74$, $p < .001$. Which dimension was relevant did not affect overall mean accuracy, adjusted Welch two sample $t(34.74) = 0.46$, $p = .65$. Most further analyses will drop dimension as a factor.

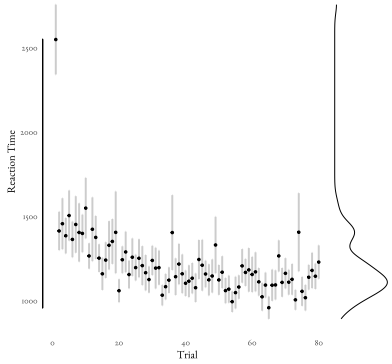

Figure 7: RFC Mean reaction time per trial during the training phase.

As the Figure 8 demonstrates, one measure important to interpreting the results of this experimental design is the distance between the trough and a given feature value, or the *trough distance*. When a stimulus had relevant feature values near the trough, which was also the category boundary, accuracy went down the nearer that feature value approached the trough. This was not the case for incidental feature values, which were unrelated to trough distance. Furthermore, the relationship between accuracy and trough distance seemed logarithmic, where

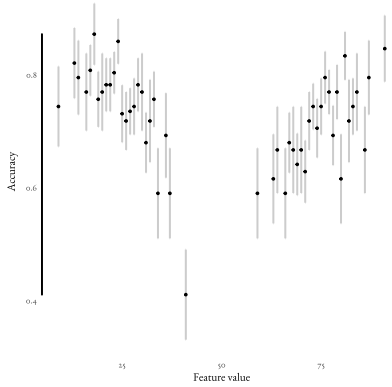

Figure 8: Experiment 1 RFC accuracy by relevant dimension feature value. Accuracy decreased closer to the category boundary.

accuracy plateaued around the feature values associated with the modes, or a trough distance of 16. The relationship between accuracy and trough distance also did not seem to vary by which of the two stimuli dimensions was the relevant one.

A generalized linear mixed model with accuracy as a dependent variable and relevant feature trough distance (log transformed), incidental feature trough distance (log transformed), relevant dimension, and the interaction between relevant and incidental trough distance as independent variables found main effects of relevant boundary distance, $\chi^2 = 35.94$, $p < .0001$, and incidental trough distance, $\chi^2 = 4.35$, $p = .04$. These effects also held without the log-transformation on relevant and incidental trough distance. No other effects or interactions contributed to model fit. As the relevant trough distance increased so did the likelihood of an accurate answer, but this effect was in the opposite direction for incidental trough distance, though seemingly a much weaker effect, and possibly due to the trail-off in accuracy for items with very high incidental feature values, as illustrated in Figure 11.

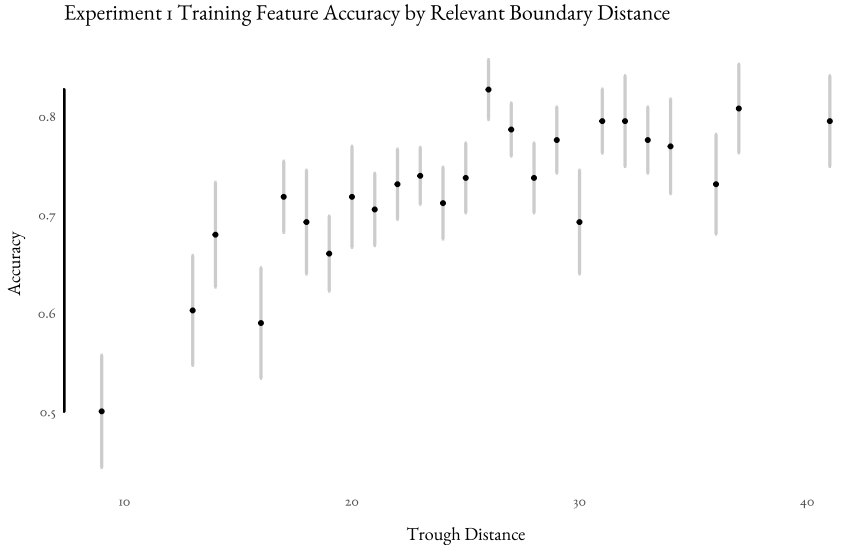

Figure 9: Experiment 1 training accuracy by relevant boundary distance. Accuracy increased with the distance from the relevant dimension's category boundary, which was in the middle of the trough of the bimodal distribution.

Figure 10: Experiment 1 training reaction time by relevant boundary distance. Reaction time was unrelated to distance from the category boundary.

Reaction time, on the other hand, did not vary with trough distance or relevant dimension in any meaningful way. A similar GLMM with reaction time as the dependent variable and the same independent variables found no main effects or interactions.

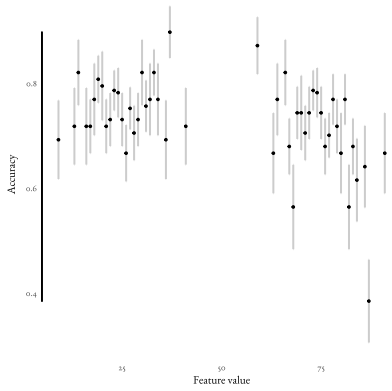

Figure 11: Experiment 1 RFC accuracy by incidental dimension feature value. Accuracy was uncorrelated with the trough.

RFC Frequency Rating Task. Using the `brms` package in R (Bürkner, 2017), we performed a Bayesian cumulative ordinal regression using rating as the dependent variable (as an ordered factor from 1 to 7), with the label for the relevant dimension as one independent fixed factor and the label for the incidental dimension as the other fixed factor, and with random intercepts for each participant. We use Cauchy priors for each of the factors which did not change the outcome of the regression compared to the default flat priors but does allow for evidence ratios for and against specific hypotheses in later comparisons across levels of different parameters. We set "mode" as our reference level for both parameters and used treatment contrasts. Using the Leave One Out Information Criterion as a measure of model fit (Vehtari, Gelman, & Gabry, 2017), the full model (LOOIC = 5027,

SE = 39.87) fit just as well as a model with no interaction between the fixed factors (LOOIC = 5015.38, SE = 39.68), both of which fit the data better than a null model or models with single fixed factors; both the relevant and incidental dimensions contribute to model fit, but their interaction does not. The following analysis will explore contrasts in the model without interactions.

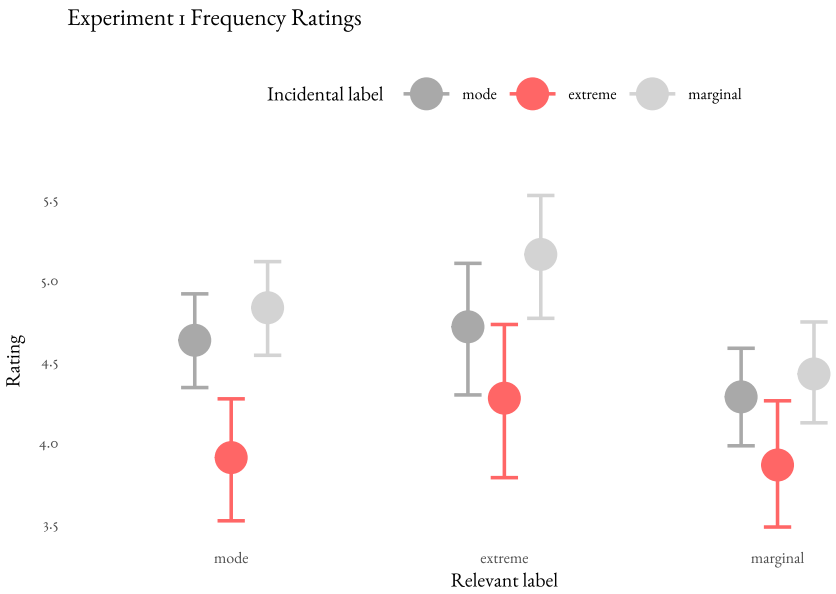

Figure 12: Experiment 1 RFC frequency rating means and confidence intervals. This figure is misleading and treats rating as continuous, but it illustrates the relatively lower ratings of the extreme incidental label.

First, the value of the incidental feature mattered for model fit, and thus for the frequency rating of the stimuli. This is evidence that people learned *something* about the incidental feature values and associated them with the frequency of a given stimulus. Second, we tested our initial hypothesis: if participants learned about the distributional structure of the feature dimensions then frequency ratings for stimuli that have modal feature values should be higher than when they have marginal or extreme feature values. Along the relevant dimension modal values were not uniformly higher than extreme, β = 0.30, $LCI_{95\%}$ = -0.02, $HCI_{95\%}$ = 0.63, or marginal feature values, β = -0.30, $LCI_{95\%}$ = -0.61, $HCI_{95\%}$ = 0.00; nor was it the case that in this most optimistic initial hypothesis that such learning happened along the incidental dimension favoring the modal over

extreme, β = -0.57, $LCI_{95\%}$ = -0.91, $HCI_{95\%}$ = -0.23, or marginal values, β = 0.31, $LCI_{95\%}$ = 0.06, $HCI_{95\%}$ = 0.55. Indeed, the only label values that were rated as less frequent than the ratings when the feature values were modal are relevant-marginal (36.38 times more likely to be rated below the relevant-modal value than above it) and incidental-extreme (two thousand times more likely to be rated below the incidental-modal value than above it).

The other hypotheses put forward in the predicted results section come down to marginal values being either highly represented and thus rated as more frequent, or as the least frequent. Along the relevant dimension they would be rated the most frequent if participants were focusing on minor details determining category membership along the relevant dimension (e.g., paying more attention when feature values were nearest the category boundary). Along the incidental dimension, they would be rated lowest as there is no category boundary with which to separate them; only extreme and modal values would be remembered. We can test this using `brms`'s `hypothesis` function. In both of these cases the intuition was wrong: relevant-marginal values were rated as less frequent than relevant-extreme, β = -0.60, $LCI_{95\%}$ = -1.03, $HCI_{95\%}$ = -0.15, and incidental-marginal values were rated as more frequent than incidental-extreme, β = 0.88, $LCI_{95\%}$ = 0.55, $HCI_{95\%}$ = 1.23. So we see that none of the hypotheses set out in the predicted results sections accurately characterizes this pattern of results, where relevant-marginal values are rated as less frequent than their alternatives and incidental-marginal values are rated as more frequent.

RFC Forced-Choice Task. During the forced choice task participants saw two stimuli that differed along one dimension, either the relevant or incidental dimension, one of which had its varying dimension's feature value set to a mode of the bimodal distribution seen during training (the correct answer) or

a marginal or extreme tail value. We performed a GLMM with accuracy as a dependent variable with dimension (relevant or incidental) and the incorrect feature value (marginal or extreme) as independent factors and random effects of those per participant. The dimension upon which the question differed had no effect on model fit, $\chi^2 = 3.82$, $p = .06$, with participants' accuracy on questions in which the incidental dimension differs the same as when the relevant dimension differed. When the incorrect feature value provided in contrast to the modal value in the forced choice questions was an extreme value, answers were more accurate than when the alternative was a stimulus with a marginal feature value, $\chi^2 = 4.55$, $p = .03$. The interaction between incorrect feature value and the varied dimension did not contribute to model fit. A similar gaussian model examining reaction time showed no contribution to model fit from either factor or their interaction; reaction time was the same regardless of the makeup of the presented stimulus.

In this sense, there is some evidence that participants developed some kind of representation, though that evidence is not particularly strong. Accuracy seems to improve when the alternative feature value is marginal rather than extreme which could be interpreted as evidence that participants were more accurate at distinguishing between modal values and marginal ones than between modal values and extreme ones; that is, they find it easier to distinguish between the center of the entire feature distribution and the most frequent parts of that distribution than between the extreme parts of the distribution (which had the same frequency as the marginal parts) and the modes.

Because of the binary response, the task could be subject to random guessing. In this sense chance performance is hard to gauge as there were not a large number of trials per participant to get a very nuanced binomial test. A one-tailed binomial test of accuracy examining each participants' number of correct responses to

a chance level of .5 with a 95% confidence interval showed that 36 of the 39 participants were at chance. This is a somewhat crude measure; out of 16 questions, in order to be considered above chance in this test, one can only get 4 questions wrong. While it is possible that people were actually biased and believed that, say, a marginal value was more frequent than a modal value, only one participant was *less* accurate than chance.

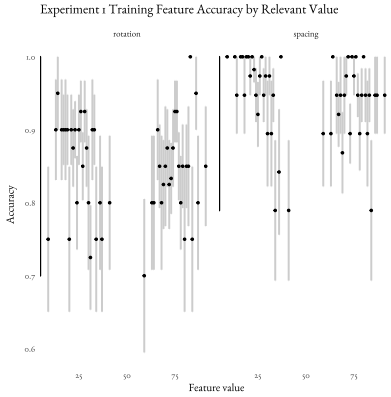

Figure 13: Experiment 1 Gabor accuracy by relevant dimension feature value.

Gabor Training. Which dimension was relevant *did* affect overall mean accuracy, adjusted Welch two sample $t(21.47) = -3.18$, $p = .005$. Which dimension was relevant will be included in the following analyses. Participants were very accurate during training. Calculating an average accuracy score for each participant, we see $\mathcal{M}_{\text{m-rotation}} = 0.85$, $SE_{\text{m-rotation}} = 0.04$, and $\mathcal{M}_{\text{m-spacing}} = 0.95$, $SE_{\text{m-spacing}} = 0.02$. Overall mean accuracy was above chance in a one sample t-test, $t(38) = 24.03$, $p < .0001$.

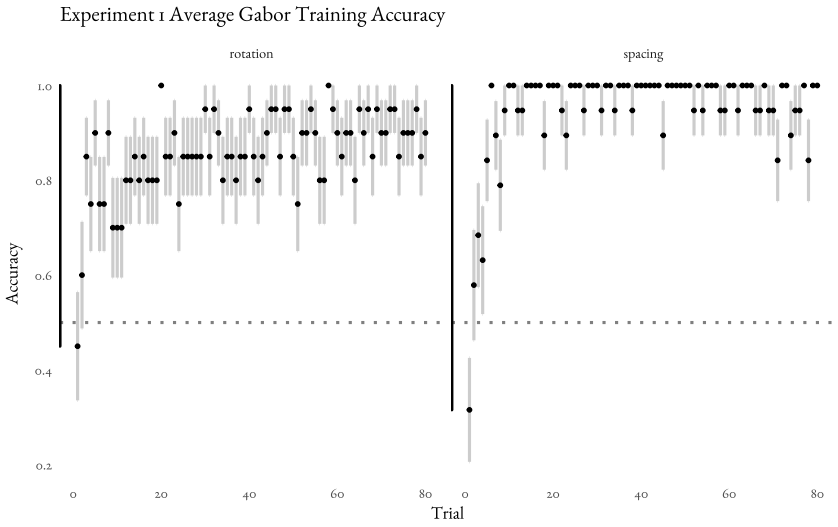

Figure 14: Accuracy during the Gabor training phase of Experiment 1. Black points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance accuracy for each trial. Accuracy was quickly at ceiling when spacing was the relevant dimension, but not so when rotation was relevant.

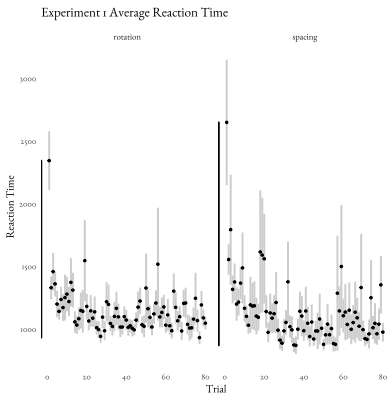

Figure 15: Reaction time during the Gabor training phase of Experiment 1. Black points represent mean reaction time for all participants for a trial. Lines represent standard error of the mean for each trial.

We performed a generalized linear mixed model with accuracy as the dependent variable modeled binomially and with the relevant dimension (rotation or spacing) as a fixed factor, trough distance of the relevant dimension, and the

trough distance of the incidental dimension as fixed parameters. This model did not converge. However, when we removed the complex random effects structure involving the interaction between relevant and incidental trough distance parameters, we find that only which dimension was relevant influenced accuracy, χ^2 = 11.85, p = .0006; no other parameter's inclusion influenced model fit. To check this, we performed a Bayesian GLMM with a similar model structure as the one above using the `brms` package in R (Bürkner, 2017). No model fit more or less accurately than the full model [LOOIC = 1770.47, SE = 71.18, Vehtari et al. (2017); WAIC = 1770.09]. The only large parameter estimate was for which dimension was relevant, β = 2.78, SE = 2.01, LCI = -1.17, HCI = 6.74. Testing the hypothesis that when the relevant dimension was spacing, accuracy was higher than when it was rotation, we find support for this one-sided hypothesis to be about twelve times more likely than the alternative (β = 2.78, SE = 2.01, LCI = -0.49). Absolute values for the parameter estimates for the relevant and incidental trough distances were both below 0.1.

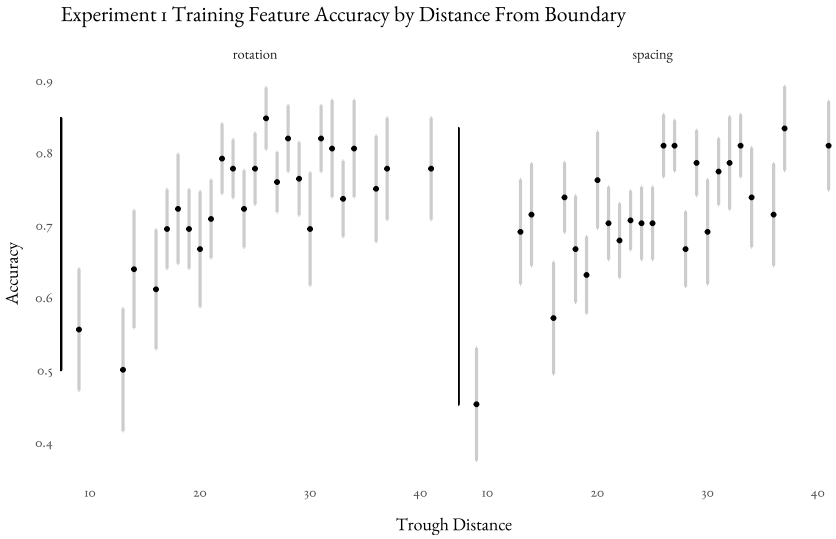

Figure 16: Experiment 1 Gabor training accuracy by relevant boundary distance.

We performed a Bayesian GLMM similar to the one above but using reaction time as the dependent variable. Which dimension was relevant had no effect on

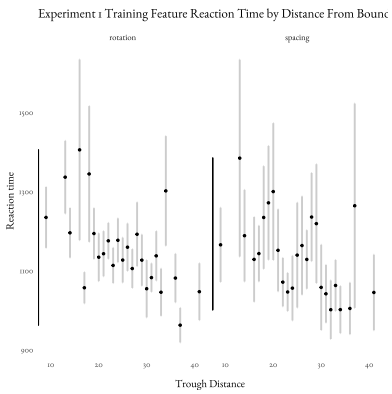

Figure 17: Experiment 1 Gabor training reaction time by relevant boundary distance.

reaction time, β = 306.64, SE = 362.49, LCI = -393.43, HCI = 1020.34, with an evidence ratio of 3.98 favoring the hypothesis that reaction time for participants trained on spacing was higher rather than reaction time for participants trained on rotation over the reverse, which is barely anecdotal and not convincing in this context. The full model fit the data best, though all permutations were within the standard error of one another, providing even more evidence that no parameter had any notable effect on reaction time.

It seems as though people who were trained with spacing as the relevant dimension were more accurate than people trained with rotation as the relevant dimension, and we did not see the effect of trough distance that we saw in participants trained on RFCs.

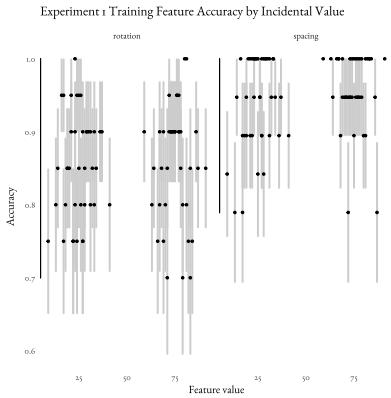

Figure 18: Experiment 1 Gabor accuracy by incidental dimension feature value.

Gabor Frequency Rating Task. As before, we performed a Bayesian cumulative ordinal regression using rating as the dependent variable (as an ordered factor from 1 to 7), with the label for the relevant dimension as an independent fixed factor, the label for the incidental dimension as a fixed factor, the relevant dimension (rotation or spacing) as a fixed factor, and with random intercepts for each participant. We use Cauchy priors for each of the factors which did not change the outcome of the regression compared to the default flat priors but does allow for evidence ratios for and against specific hypotheses in later comparisons across levels of different parameters. Again, we set "mode" as our reference level for both parameters and used treatment contrasts. Using the Leave One Out Information Criterion as a measure of model fit (Vehtari et al., 2017), the full model (LOOIC = 4806, SE = 49.08) fit just as well as a model with no interaction between the fixed factors (LOOIC = 4791, SE = 48.47), both of which fit the data better than a null model or models with single fixed factors; both the relevant feature dimensions contribute to model fit, but their interaction does not.

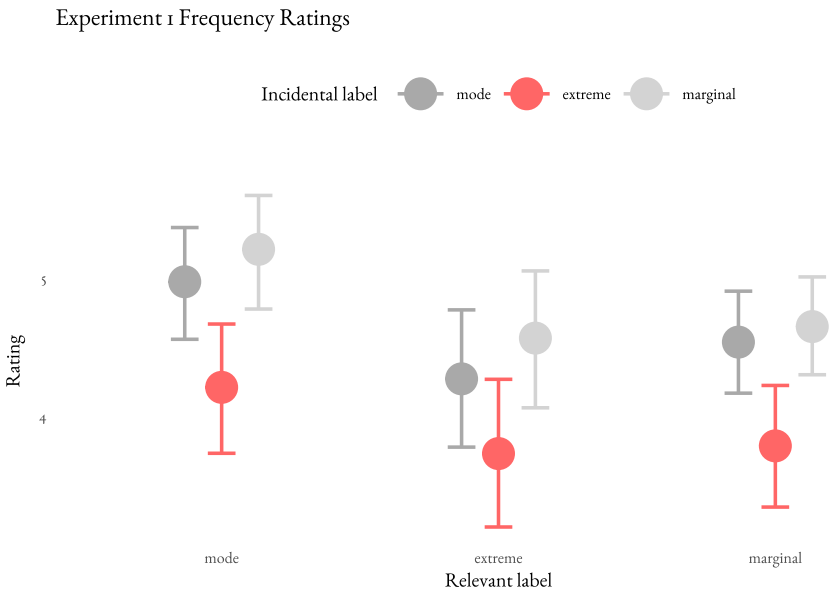

Figure 19: Experiment 1 Gabor frequency rating means and confidence intervals. This figure is misleading and treats rating as continuous.

Gabor Forced-Choice Task. As with the RFC stimuli, we performed a GLMM with accuracy as a dependent variable, dimension (relevant or incidental) and incorrect feature value (marginal or extreme, always presented in in contrast to the correct answer, one of the two modal values) as independent factors, and random effects of those per participant. Training dimension was also included as a factor because it seems to matter for the Gabor stimuli, as seen in the training section. Whether the question presented differences in the relevant or incidental dimension did not contribute to model fit, $\chi^2 = 1.08$, $p = .30$; accuracy was the same regardless of the dimension that differed. Like in the previous experiment with RFC stimuli, participants were more accurate when the incorrect feature presented in contrast to the correct modal feature was extreme than when it was marginal, $\chi^2 = 5.28$, $p = .02$. The interaction between dimension and the incorrect feature value did not contribute to model fit. However, there was a three-way interaction between which dimension was relevant during training, which dimension differed during the forced choice task, and the label for the correct option, $\chi^2 = 3.88$, $p = .05$. Interpreting this three-way effect is difficult,

and a Tukey pairwise comparison of all the levels within the interaction gives no differences after adjusting the p-values using the Tukey method.

A similar model with reaction time as the dependent variable indicated that only the incorrect feature value contributed to model fit, $\chi^2 = 5.98$, $p = .01$, with no effect of dimension or of the interaction between dimension and the incorrect feature value. Reaction time was faster when the incorrect feature value presented in contrast to the modal feature was marginal than when the incorrect feature value was extreme. There was also a main effect of the dimension that was relevant during training, $\chi^2 = 4.46$, $p = .03$, such that responses from people who went through training with the relevant dimension being rotation were faster (M = 1690ms) than those who were trained with spacing as the relevant dimension (M = 2051ms). No other main effects or interactions were present.

In a forced-choice task, when presented with a correct modal feature value, accuracy increased when the alternative choice was extreme rather than marginal. This result is the opposite of the same task when performed with RFC stimuli. Furthermore, the perceptually separable Gabor stimuli show an effect on reaction time, such that participants are faster when the alternative choice is marginal rather than extreme. This is in the opposite direction as would be expected (if there were any expectation for reaction time at all).

Again, as was said in the RFC Forced Choice Task results, because of the binary response the task could be subject to random guessing. A one-tailed binomial test of accuracy examining each participants' number of correct responses to a chance level of .5 with a 95% confidence interval showed that 34 of the 39 participants were at chance. Again, this is a crude measure; out of 16 questions, in order to be considered above chance in this test, one can only get 4 questions wrong.

Discussion

Experiment 1 was a success for the purposes of this dissertation. We demonstrated that participants engaged in a supervised classification task were able to learn about the relevant dimension within the span of the training phase, and learned more about the relevant dimension than the incidental dimension. This demonstrates the validity of our version of the supervised classification task, which was the primary goal of Experiment 1. Furthermore, the experiment confirmed our suspicions about the stimuli sets we tested. Participants trained on RFCs learned during training and showed no bias toward one dimension over the other, while participants trained on Gabor patches we more or less accurate and learned more or less depending on which dimension was presented as relevant during the training phase. This adds evidence to our suspicion that with Gabor patches there will be less learning about the incidental feature dimension in a supervised classification task. While it would be possible to balance our Gabor stimuli dimensions to compensate for any biases in them analytically, this would take us down a path tangential to the purpose of this dissertation. Given the sizable literature on learned inattention in separable dimensions and the relative lack of such work on integral stimuli, this gives us a better chance to explore a less charted area of the supervised classification task, as well as potentially increases our chances to get people to learn about the distribution of the incidental feature distribution during supervised classification.

Participants clearly learned something about the incidental feature values during training, as seen in the results of the frequency rating task. However, our intuition about what relevant and incidental dimension feature values would garner higher frequency ratings was wrong. Relevant-marginal values were rated as less frequent than relevant-extreme, and incidental-marginal values were rated

as more frequent than incidental-extreme. None of the hypotheses set out in the predicted results sections accurately characterizes this pattern of results across both relevant and incidental dimensions, where relevant-marginal values are rated as less frequent than their alternatives and incidental-marginal values are rated as more frequent. But some combination therein accounts for the pattern of results: It could be that extreme values are rated as more frequent along the relevant dimension because participants have a kind of prototypical representation of the relevant dimension; they know the range and learn enough to distinguish between extreme high and low values, which have a definitive category membership because of their distance from the category boundary, whereas marginal values along the relevant dimension are not represented or remembered as well because of their proximity to the category boundary. Along the incidental dimension, however, the opposite seems to be the case. Incidental-marginal values were rated as more frequent than other values. Why? One possibility discussed during the predicted results section is that because the dimension is not diagnostic but does vary continuously it could garner a kind of representation as a normal distribution, where more central values (e.g., marginal values) are represented as more frequent, and more extreme values are represented as less frequent. That is, participants might have learned something about the distributional properties of the incidental dimension, in that it *has* a distribution, but may not have learned that its distribution was bimodal, or anything else *about* that distribution. More generally, while participants in Experiment 1 are learning about the incidental dimension, they may not be learning two clusters with a trough between them, but learning the feature distribution as one large continuous cluster.

Evidence for anything in the forced choice paradigm was limited; it is possible many people were at chance. If their results were biased because of their represen-

tation of the feature space and not because of guessing, the only effect was that they were more likely to pick the correct modal value over the incorrect extreme value than over an incorrect marginal value.

Taken together, the results of Experiment 1 give evidence that, when the dimensions are integrated (e.g., when the stimuli are radial frequency components), participants in a supervised classification task overrepresent the frequency of marginal feature values in the incidental dimension and underrepresent them along the relevant dimension. This, at least, shows evidence that people *do* learn about the incidental feature dimension, and *do* represent it, but in a way that is different than the way they represent the relevant feature dimension.

Experiment 2

The previous experiment confirmed that people do learn the relevant dimension of our RFC stimuli in a supervised classification task, and asked whether participants learned about the incidental feature dimension when learning a category boundary along the relevant dimension. Experiment 2 asks whether or not information in the incidental feature dimension bears on accuracy when the supervised classification task changes such that the incidental feature subsequently becomes relevant. The purpose of Experiment 2 is to determine what people learn about incidental feature information in a standard categorization task when the feature distributions have statistical properties that could potentially help (in the case of a similar or consistent category boundary) or harm (in the case of an incongruous or inconsistent category boundary) learning new categories involving the previously-incidental feature. The way we do this to manipulate the category boundary in the second supervised classification phase to be either consistent or inconsistent with the location of the trough in the first phase's incidental feature distribution.

In Experiment 2, participants categorized stimuli in a training phase similar to Experiment 1, where both the t1-relevant and t1-incidental stimuli features are uncorrelated and distributed in a bimodal distribution. In this and the following experiments "t1" refers to the first of two tasks, sometimes called the "training phase," and "t2" refers to the second. While the language is somewhat clunky,

it is important to be descriptive of what dimension is relevant when. (It follows that the t1-relevant feature becomes the t2-incidental feature, and the t1-incidental feature becomes the t2-relevant feature.) The trough of the t1-relevant feature dimension's bimodal distribution marks the category boundary. Participants then perform another supervised classification task with the stimuli distributed in a new way. They switch to learning a different set of categories based on the previously t1-incidental, t2-relevant feature. In this new t2 phase where both relevant and incidental features are uniformly distributed. Participants were assigned to one of two conditions during t2, either a *consistent* condition where the category boundary is at the trough of the incidental feature distribution presented during t1, or an *inconsistent* condition where the category boundary was at the same feature value as one of the modes in t1.

If participants learned about the incidental feature distribution then the consistent group should be more accurate than the inconsistent group during t2. The inconsistent condition's category boundary during t2 is at the value of what was a peak in the feature distribution during t1. Do people learn something about the distribution of task-irrelevant features? Does the category boundary in the inconsistent condition, in a different place than t1 during t2, hurt accuracy, even though the relevant feature dimension during t2 was irrelevant during the t1 phase? Another way of saying this is that in the inconsistent condition, the t2-relevant category boundary could split the cluster structure that participants might have learned about the t1-incidental feature distribution during the t1 phase.

Experiment 2 was also a chance to see if we could get away with paying less to participants for completing the Experiment. If MTurkers are just as accurate when paid $1.50 for completing the experiment as those previously paid $3.00, then it would make sense to pay half as much and recruit twice as many participants. As

we will see, participants paid $1.50 for their time were less accurate overall and more likely to give responses at chance accuracy than those paid $3.00. As such, Experiment 2 should be seen as exploratory in nature, a kind of first pancake of the two-phase paradigm the rest of the dissertation uses. Experiments 3–5 all reward $3.00.

Methods

Participants. A total of 100 participants completed Experiment 2. No participants participated in more than one experiment. 79 participants completed the task with a $1.50 and 21 completed the task with a $3.00 reward as a test of whether reward amount influenced t1 accuracy, which it did, as we shall see later.

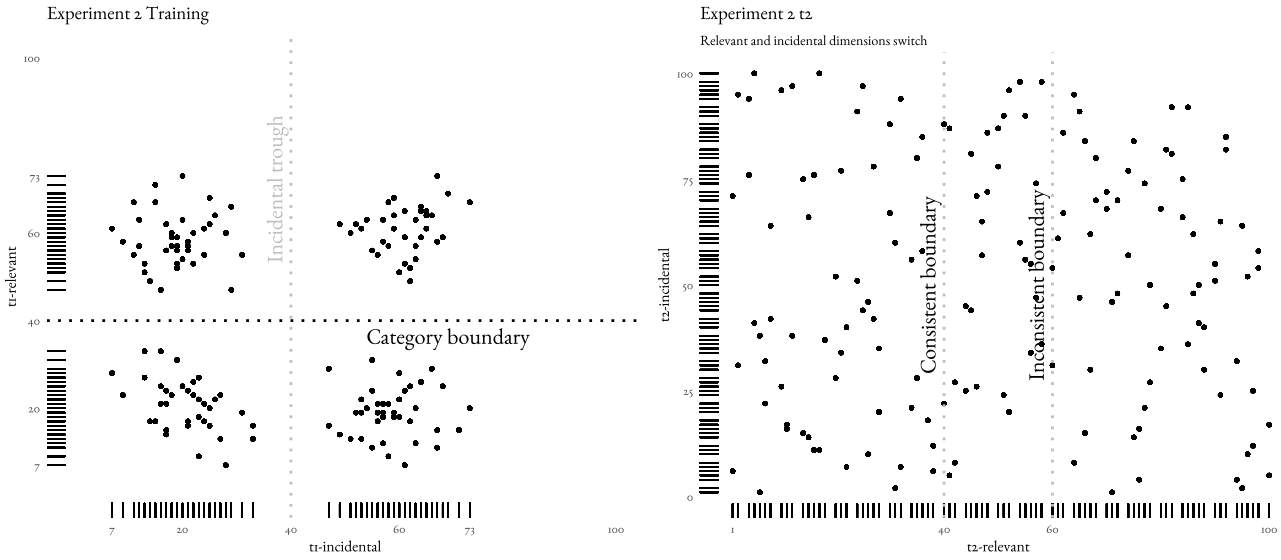

Figure 20: Experiment 2's feature distributions and category boundaries in the training and testing phases.

\

Design. The experiment consisted of two phases, the training phase, t1, and the testing phase, t2. Participants were randomly assigned to either the inconsistent condition or the consistent condition before the beginning of the experiment.

These conditions reflected whether or not the category boundary during the t2 phase is consistent with the trough in the incidental feature distribution during the training phase.

The design of t1 was similar to the training phase in Experiment 1 but with different values and a category boundary at 40, the trough of the t1-relevant feature distribution. Around that category boundary trough were centered two modes at 20 and 60 making a symmetrical bimodal distribution for each feature dimension, which are uncorrelated with one another. Participants categorized 80 stimuli in the t1 phase into "feps" and "blickets" and received feedback after each classification.

The t2 phase tasked participants to learn a new category boundary between two new categories, "wugs" and "plits." Participants categorized 80 more stimuli in the t2 phase and received feedback after each classification. The stimuli were very similar to the previous phase in all ways except their distribution and category membership. These categories' relevant feature was the incidental feature in t1. All stimuli presented during t2 were drawn from a uniform distribution from 1 to 100 for each dimension, which were uncorrelated with one another. In the consistent condition, a "wug" was an RFC stimulus where the t2-relevant feature has a value of 40 or less; that is, the category boundary was consistent with the trough in the t1-incidental feature distribution seen during t1. Everything that was not a wug was a plit. In the inconsistent condition, the category boundary was 60, where the high mode of the t1-incidental feature distribution was. The consistent condition had a 40%/60% split of wugs to plits, while the inconsistent condition had a 60%/40% split.

Materials. The same set of radial frequency component stimuli from Experiment 1 were used in Experiments 2, 3, and 4. The material differences between the

Experiments resulted from the distribution of the stimuli during both phases as described in the Design section. In the t1 phase, the stimuli features represented only 80% of the available feature space seen during t2. Both dimensions present in the t1 stimuli were distributed bimodally with modes at feature values of 20 and 60 with a standard deviation of 6 about them. During t2 the feature distributions were uniform.

Procedure. The instructions for Experiment 2's t1 phase are nearly the same as Experiment 1, except for one difference. Rather than asking participants to categorize stimuli into "feps" and "non-feps," the instructions directed participants to categorize them into two labeled categories: "feps" and "blickets." For the t2 phase, the same instructions were presented, with "wug" substituted for "fep" and "plit" for "blicket."

Results

Metrics. This section outlines a number of metrics that will be examined in this and the following experiments, which share a similar methodological and procedural structure.

The goal of Experiment 2 is to see whether previously incidental information influences learning and accuracy in supervised classification when it becomes relevant. The simplest method of analysis is to look for the main effect of consistency in t2. If the inconsistent condition is sufficiently thrown off by the t2-relevant category boundary being different from the location of the t1-incidental trough, then participants in the inconsistent condition should be less accurate in their classifications than participants in the consistent condition. This method of analysis is blunt and lacks nuance, but would do just fine if the effect size is large. Pilot data

indicated it is not.

One metric to examine in Experiment 2 is to calculate a gain score. What is the difference between accuracy for each participant between t1 and t2, then analyze that by the factor of consistency. One hypothesis is that participants who encounter the consistent t2 phase will have a higher gain score than those in the inconsistent condition. The relative gain score also allows us to retain all data as it roughly controls for individual differences in learning rate or how engaged each participant was. Someone less engaged or distracted in the task might have a lower score, but show the same *difference* between t1 and t2 scores as someone who was vigilant.

Another useful metric is to calculate a regression based on the distance of the t2 stimuli from the trough present in t1. If we were to calculate a regression using accuracy as a dependent variable and boundary distance (the distance from the t2-relevant feature and the category boundary during t2), as well as condition, as independent variables, a hypothetical participant coming into the t2 phase fresh, without first encountering the t1 phase would have trouble discriminating stimuli near the category boundary, and find it easier as the relevant feature value moved away from that boundary. This serves as a null hypothesis: performance in the t2 phase is just a feature of how far each stimulus is from the category boundary. If this is the case, there should be no difference between participants in the consistent and inconsistent conditions.

As the Overview of Experiments alluded to, the key metric in these Experiments is that of the conflict item. To restate, the main hypothesis these experiments seek to test is that in supervised classification with two dimensions the learner does unsupervised learning of the incidental feature during t1. In Experiments 2 through 5, there are two clusters with a trough in the middle—a cluster

with low feature values and a cluster with high feature values. The learner can represent that as a cluster structure. Are boundaries that respect this cluster structure easier to learn than boundaries that violate that cluster structure in the t2 phase? What it means to violate that cluster structure, to be misaligned with it, is the case when stimuli with relevant feature values in the same cluster end up on different sides of the category boundary. In unsupervised learning of the t1-incidental feature distribution during the first task, there are no labels so the dimension could be thought of as one large cluster (a potentially Gaussian one in our unimodal hypothesis). But in the supervised learning of the t2-relevant feature during the second task, the clusters are defined by labels. The label and the distribution conflict for some items. These are called "conflict items."

Conflict items are items presented in the inconsistent condition in t2 that have relevant feature values that conflict with their t1-incidental cluster structure. More specifically these are the stimuli with relevant feature values between 40 and 59. Stimuli presented during t1 with t1-incidental feature values in the range between 40 and 59 constituted about half (the left half, to be visual) of a normal distribution with a mode at 60. The category boundary for participants in the inconsistent condition during t2 is also at 60. Conflict items are the stimuli with relevant feature values that should be the most difficult for people in the inconsistent condition to classify because of the discrepancy between their dimension's distribution during t1 and the location of the category boundary during t2. If any intuitions are built up in the t1 phase about the incidental feature distribution, these conflict items go against those intuitions, but only in the inconsistent condition.

We can contrast differences in performance on the different sides of the conflict items' category boundary with performance on similar sides of the category

boundary in the consistent condition, equivalent to 20-39 and 40-60. This serves as a baseline comparison between conditions. In the consistent condition, there should never be any difference in accuracy on either side of the category boundary because there is no shift from t1 to t2, or even in the feature values between dimensions, that would contribute to such a difference. If there was a difference in accuracy or reaction time we would know something was awry with the way participants were interpreting our stimuli.

There are other things people could learn in the t1 phase that they bring to the t2 phase. If participants represent the t1-incidental dimension as a single cluster, as a kind of uniform or normally-distributed feature dimension, as discussed in Experiment 1, one strategy they could use in the t2 phase of Experiment 2 is to divide the stimuli into two groups: the cluster they learned about during t1, and anything else. In this sense, the cluster with which they represent the incidental dimension is assigned one category label and all new values to the other label. In the t2 phase of Experiments 2 and 3, both stimuli feature dimensions are uniform and extend from 1 to 100. While the consistent condition has a category boundary at 40, consistent with the trough of the t1-incidental feature, the inconsistent t2-relevant feature has a category boundary at 60, where a mode was in the t1 phase. This means that one strategy for accurate performance in t2 for people in the inconsistent condition is to use novelty as a predictor—rather than focus on the category boundary at 60, this strategy effectively shifts the category boundary to 73, which is not a large shift in the number of potentially miscategorized stimuli. If a participant adopts this strategy, their task is to decide whether or not they've seen a given stimulus. If they have, they categorize it as a "wug." If they haven't, they categorize it as a "plit." That is, the categories become familiar stimuli and unfamiliar stimuli, the cluster they were exposed to during t1 and the new cluster,

even though in t2 neither of these features have a distinct cluster structure. People in the inconsistent condition who adopt this strategy would do poorly on stimuli with a relevant feature value between 60 and 73 (or some reasonable threshold of novelty), which make up 11 trials. But people in the consistent condition could not adopt this novelty strategy and be successful; because their critical but familiar items would range from 40 to 73, fully half of the familiar items they classify during t1 would be miscategorized. If participants in the consistent condition (or in both conditions) adopted this strategy and used this representation, then accuracy for people in the inconsistent condition would be higher than those in the consistent condition. However, accuracy alone is not a fully descriptive method of analysis; even if participants in the inconsistent condition learned the t1-incidental, t2-relevant distribution perfectly, based on the results of Experiment 1 we would expect accuracy around a condition's category boundary to be lower the closer stimuli were to that boundary, so even if both stimuli dimensions were represented continuously, we would expect accuracy to be lower for the inconsistent condition on items between 60 and 73 as they border the inconsistent condition's category boundary in t2.

Finally, one concern about examining all 80 t2 trials is that any effect of the t1-incidental distribution could be washed out by efficient learning of the t2 task alone. Each analysis section will look at the first third (24 items) of t2 trials to see if there is any indication of an effect of the consistency condition early on in the t2 phase.

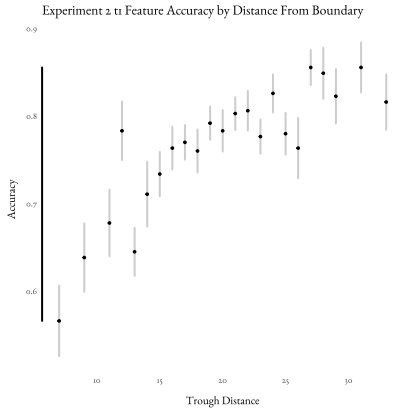

Figure 21: Experiment 2 training accuracy by relevant boundary distance.

T1

Participants were somewhat accurate during t1. Calculating an average accuracy score for each participant, we see $M = 0.71$. Participants who received the $1.50 reward had an overall t1 accuracy of $M = 0.69$, while participants who received the

$3.00 reward showed $\mathcal{M}$ = 0.76, $t(33.89)$ = -2.06, p = .048, however BF_{10} = 0.46. Overall, experiment reward did not affect accuracy during t1. While the difference is significant in a normal t-test, the Bayes Factor is small enough to cast doubt on that result. Either one of these statistics could be due to the low number of participants in the $3.00 version.

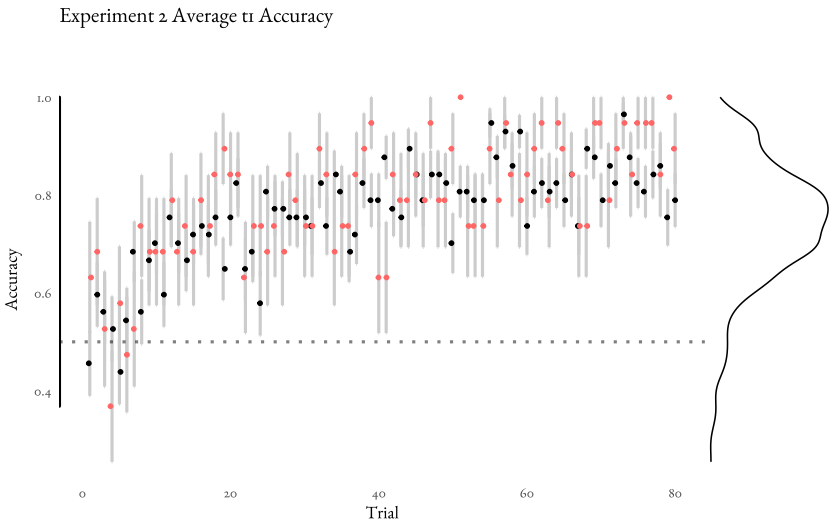

Figure 22: Accuracy during the t1 phase of Experiment 2. Black points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance accuracy for each trial. Red dots indicate higher payout.

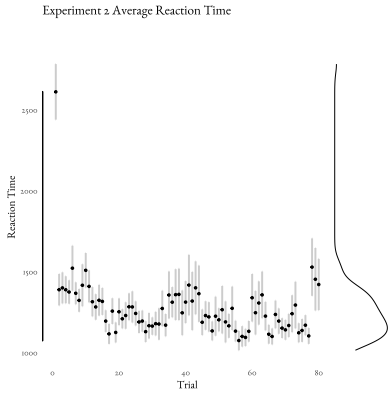

Figure 23: Reaction time during the t1 phase of Experiment 2. Black points represent mean reaction time for all participants for a trial. Lines represent standard error of the mean for each trial.

However, because of the large number of participants that performed at chance during t1 in the $1.50 reward condition, this analysis will only look at those participants whose accuracy was above chance during t1. Twenty-two participants who received $1.50 for participation were at or below chance accuracy, while only 2 were at or below chance who received a $3.00 reward. That leaves 57 participants in the $1.50 reward condition and 19 participants in the $3.00 reward condition that will be reported on together in the following sections. Together, participants' mean accuracy was $\mathcal{M}_{\text{accuracy}}$ = .77, with $\mathcal{M}_{\text{RT}}$ = 1272ms. To test whether there were any *a priori* group differences by condition (which does not vary in t2), a GLMM with t1 accuracy as the dependent variable and consistency and reward amount as independent variables found no effects or interactions. There were no *a priori* group differences in t1 accuracy.

T2

Overall Accuracy. A logistic generalized linear mixed model with accuracy as a dependent variable and consistency as an independent fixed variable and participant and trial number as random factors showed no influence of consistency on model fit compared to the null model, $\chi^2 = 0.33$, $p = .56$. A similar repeated measures ANOVA also showed no influence of consistency on accuracy, $F(1, 74) = 0.36$, $p = .55$. However, the same was not true for a model with reaction time as the dependent variable, $\chi^2 = 4.90$, $p = .03$, $F(1, 74) = 4.93$, $p = .03$. Interestingly, people in the inconsistent condition were faster to respond than participants in the consistent condition, $\mathcal{M}_{\text{inconsistent}} = 1120$ms, $\mathcal{M}_{\text{consistent}} = 1313$ms.

Because the hypothesized effect of consistency was probably weak and could be drowned out by learning over the t2 phase, it is informative to examine the first 24 items, the first 1/3rd of the t2 phase, in order to see just how category boundary consistency affects accuracy and reaction time. In the first 24 t2 items a GLMM showed no effect of consistency on accuracy, $\chi^2 = .30$, $p = .58$, $F(1, 74) = 0.29$, $p = .59$. However, there was an effect of consistency on reaction time, $\chi^2 = 6.92$, $p = .009$, $F(1, 74) = 7.10$, $p = .009$. Again, people in the inconsistent condition were faster to respond than participants in the consistent condition, $\mathcal{M}_{\text{inconsistent}} = 1178$ms, $\mathcal{M}_{\text{consistent}} = 1448$ms.

Gain Score. Both a GLMM and a repeated measures ANOVA using gain score, measuring the difference between accuracy in t1 and t2 for each participant as a dependent variable and condition as a fixed independent variable showed no difference between t1 and t2 accuracy by condition, $\chi^2 = 0.22$, $p = .64$, $F(1, 74) = 0.21$, $p = .65$. There was also no effect of consistency on gain score on the first 24

items presented during t2.

Boundary distance. In supervised classification experiments many studies have shown that the closer a stimuli's feature value is to the category boundary the harder it is to determine its category membership (e.g., Ashby, Boynton, & Lee, 1994). This is what we predict would matter if people were starting t2 fresh without any influence from t1. One way we can control for that is by introducing a new covariate, boundary distance, or the t2-relevant, t1-incidental feature's distance from the t2 category boundary. A logistic GLMM using accuracy as the dependent variable, consistency as a fixed independent factor, boundary distance as a fixed covariate, and the participant by relevant boundary distance interaction (the within-subject variable) and trial number as random variables showed no effect of consistency on accuracy, $\chi^2 = 0.74$, $p = .39$, nor did the model comparison show an interaction between consistency and relevant stimuli boundary distance, $\chi^2 = 2.72$, $p = .10$. As would be expected, there is a strong relationship between boundary distance and accuracy, $\chi^2 = 185.56$, $p < .0001$, such that items nearer the boundary are more difficult than items further away. However, consistency does seem to affect model fit when reaction time is the dependent variable, $\chi^2 = 73.63$, $p < .0001$, while trough distance plays little role, $\chi^2 = 3.30$, $p = .07$, along with their interaction, $\chi^2 = 0.59$, $p = .44$. However, this effect was in the opposite direction predicted, with $M_{\text{consistent}} = 1312$ ms and $M_{\text{inconsistent}} = 1118$ ms. It is possible this difference in reaction time could be due to participants in the different conditions using different strategies: those in the inconsistent condition could be using a kind of faster novelty strategy.

The first 24 items of t2 yield the same pattern of results. A GLMM showed no effect of consistency on accuracy, $\chi^2 = 0.23$, $p = .63$, a main effect of trough distance, $\chi^2 = 39.60$, $p < .0001$, and no interaction, $\chi^2 = 1.29$, $p = .26$. Consistency

did affect reaction time, $\chi^2 = 41.04$, $p < .0001$, while boundary distance and their interaction did not contribute to the model, $\chi^2 = 0.01$, $p = .92$, and $\chi^2 = 0.65$, $p = .42$, respectively.

Conflict Items. A logistic GLMM within the inconsistent with accuracy as the dependent variable and conflict as the independent fixed factor (40-59 vs 60-80) with the interaction between conflict and participant as the random factor showed no effect of conflict on accuracy. That is, in the inconsistent condition there was no difference in accuracy between t1-relevant stimuli with feature values between 40 and 59 and those between 60 and 80, $\chi^2 = 1.88$, $p = .17$. The same was true for reaction time, $\chi^2 = 0.04$, $p = .84$. However, looking at the first 24 items in t2, we saw that conflict did have an effect on accuracy, $\chi^2 = 5.99$, $p = .01$, $B_{conflict} = -0.20$, $B_{non\text{-}conflict} = 0.36$. In the first 24 t2 items, participants were less accurate on conflict than non-conflict items. Within the first 24 t2 items, conflict did not influence reaction time, $\chi^2 = 2.04$, $p = .15$. In contrast, in the consistent condition there was no difference in accuracy between the items to the left of the category boundary (20 to 39) and the items to the right (40 to 60), $\chi^2 = 1.94$, $p = .16$, nor reaction time, $\chi^2 = 0.06$, $p = .8$. Similarly, there was no difference when examining accuracy for the first 24 items, $\chi^2 = 0.13$, $p = .72$, or in reaction time, $\chi^2 = 0.39$, $p = .53$. It is also worth comparing accuracy and reaction time between consistency conditions for items with feature values between 40 and 60. There was no difference in accuracy by consistency in GLMM with consistency as a fixed factor, $\chi^2 = 1.98$, $p = .16$. The same was true for reaction time, $\chi^2 = 3.27$, $p = .07$.

Discussion

Experiment 2 sought to determine what people learn about an incidental feature that they then use once that feature becomes relevant in a supervised classifica-

tion task. It is worth briefly restating the hypotheses: During t1 the t1-incidental, t2-relevant feature dimension has a cluster structure. If people notice and use that cluster structure in t2 we would expect differences between the consistency conditions—if participants see the t1-incidental distribution as two groups of objects with modes of 20 and 60 with a trough at 40—then that cluster structure either maps onto the t2-relevant supervised category boundary (consistent condition) or not (inconsistent). The null hypothesis of this study is that t1 has no influence on the t2 phase, and as such, there should be no differences in the consistency factor. That is, if participants learned nothing about the incidental feature value, then no differences in t2 performance should present themselves. Another hypothesis is that participants could learn the range, but not the parameters of the incidental feature distribution. During t1 they were exposed to stimuli with feature values ranging from 1 to 80. Upon seeing stimuli with higher feature values from 81 to 100, participants recognize these stimuli as novel. This is a kind of incidental learning. Finally, participants could learn the parameters of the incidental feature distribution, namely the locations of the modes and trough.

The main result of note is that in the first 24 items presented during t2, but not overall, people in the inconsistent condition were more accurate on non-conflict items than on conflict items. To restate the relevant part of the Metrics section, at the root of this dissertation is the hypothesis that in supervised classification with two dimensions the learner does unsupervised learning of the incidental feature during t1. In the case of Experiments 2 through 5, there are two clusters with a trough in the middle—little ones and big ones. The learner might represent that as a cluster structure. A hypothesis of this and the following experiments is that in t2 the boundaries that respect this cluster structure will be easier to learn than boundaries that violate that cluster structure. The case when stimuli with

relevant feature values in the same cluster end up on different sides of the category boundary violates that cluster structure. Conflict items are the stimuli with relevant feature values that should be the most difficult for people in the inconsistent condition to classify because of the discrepancy between their dimension's distribution during t1 and the location of the category boundary during t2. If any intuitions are built up in t1 about the incidental feature distribution, these conflict items go against those intuitions, but only in the inconsistent condition. The results of Experiment 2 indicate that these intuitions are built up, but are potentially overwritten or discarded over the course of t2, such that there is no effect of conflict over the course of t2, but only in the first 24 items presented during t2.

The results of Experiment 2 were mostly null with the exception of an effect of consistency on reaction time with or without accounting for variance in accuracy resulting from a given stimuli's relevant feature value's distance from the category boundary. Participants in the inconsistent condition were faster to respond during t2 than participants in the consistent condition. This was borne out even if only the first third of t2 responses were analyzed. As discussed previously in the Metrics section, one strategy people in the inconsistent condition could use to account for this pattern of results is to make category predictions based on novelty. The category boundary for people in the inconsistent condition was at 60, very near the end of the tail of the distribution of t1 items. Participants in the inconsistent condition could have made their categorization decision based solely on whether or not they saw a given feature value before and not suffer much in overall accuracy. Participants in the consistent condition could not have used this strategy as their category boundary was too far away from the novelty cutoff and they would thus suffer in accuracy, so they had to make a categorization decision based on the feature values themselves rather than those features' novelty.

There were many limits on the effectiveness of Experiment 2 to examine learning about the incidental feature. One issue that could contribute to weak effects is the transition from a bimodal cluster structure in t1 to a flat uniform distribution in the t2 phase. This is addressed in more detail in Experiments 4 and 5. The experiment also transitions from having categories of equal size (40 feps and 40 blickets in t1) to having a biased distribution in t2 (48 of one category and 32 of the other depending on condition). In other words, even the consistent condition was not that consistent. This is addressed in Experiment 5. But at least within the feature distributions we used within Experiment 2, there are improvements we can make to the experimental protocol before we change the distribution of the stimuli. While the reward amount did not seem to influence accuracy in the presented statistical tests, there are other changes in addition to the reward amount that can be made to the protocol to explicate the feature distribution of the t1-incidental, t2-relevant feature during t1. In the t1 phase of Experiment 2, the participants were not told anything about how the relevant dimension would change during the transition to the t2 phase. The instructions in Experiment 3 make clear that there are two categories in t1 and t2, and the dimension upon which the categories were based will change during t2. Experiment 3 and all further experiments also increase the reward to $3.00 for completing the experiment.

Experiment 3

Experiment 2 investigated how incidental information presented during supervised classification influenced learning once that information became relevant. Experiment 3 seeks to build on the lessons learned in Experiment 2 and attempts to make more explicit to the participants the relationships between the dimensions and the phases of the Experiment. Participants are paid $3.00 to complete the same task described in Experiment 2, the only difference being the explicit instruction given before the t1 and t2 phases. In Experiment 3 the participants were told that their job is to categorize the stimuli into two categories. They are additionally told that the stimuli have two dimensions and that one dimension is relevant to their categorization task while one is incidental. Instruction before t2 reiterates this and declares that in the t2 phase the participants should use the other previously incidental dimension to categorize the similar stimuli into new categories. Note that the radial frequency component stimuli (RFC) feature dimensions are integrated and identifying just what a "dimension" is is a non-trivial task.

Methods

Participants. A total of 40 people recruited from Amazon Mechanical Turk completed the experiment. No one who attempted Experiments 1 or 2 participated

in Experiment 3.

Design. The design, materials, and procedure of Experiment 3 were exactly the same as Experiment 2, with one critical difference. The t1 instructions for Experiment 2 explicitly stated the presence of two categories, "feps" and "blickets." Experiment 3 goes further and instructed the participant that those categories were dependent on one of two dimensions present in every stimulus. The instructions for the t2 phase were similar, stating that the relevant dimension is going to switch, and the new categories of "wug" and "plit" are now based on this new and previously incidental dimension.

Materials. The materials used in Experiment 3 are the same materials used in Experiment 2.

Procedure. The only difference between the procedures of Experiments 2 and 3 is the inclusion of the following text in the instructions before the t1 and t2 sections of Experiment 3.

> These pictures vary in two different dimensions. The pictures can be classified in two different ways: as either Feps or Blickets, or as either Wugs or Plits. In the first part of the experiment, your task is to determine whether or each of those pictures is a Fep or a Blicket.

And during t2:

> Remember, these pictures vary in two different dimensions. The pictures can be classified in two different ways: as either Feps or Blickets, or as either Wugs or Plits. In this part of the experiment you are going to see similar pictures. Your task is to determine whether or not each of those pictures is a Wug or a Plit.

Participants were rewarded with $3.00 for completing the task.

Results

Experiment 3 examines the same metrics mentioned in Experiment 2.

T1

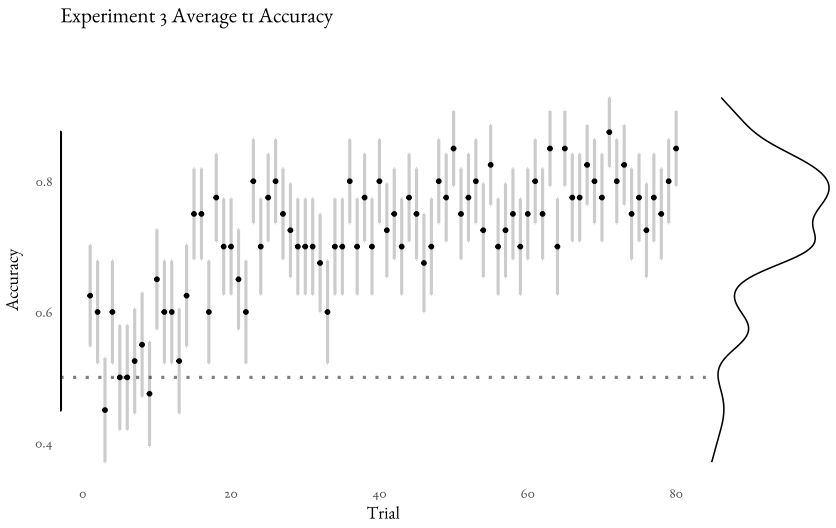

Figure 24: Accuracy during the t1 phase of Experiment 3. Black points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance accuracy for each trial.

Again, which dimension was relevant during t1 did not affect overall accuracy for participants trained on RFCs, Welch's $t(37.58) = 0.37$, $p = .716$, $BF_{10} = 0.12$. All further analysis will drop dimensionality as a factor.

Most people learned during t1. Only five participants performed at or below chance. Calculating an average accuracy score for each participant, we see $M_{acc} = 0.72$, $\mathcal{M}_{RT} = 1184$ms. Accuracy in participants later assigned to the consistent condition was $M_{acc} = 0.73$, and accuracy in those later assigned to the inconsistent condition was $M_{acc} = 0.70$. As would be expected, because the manipulation is not yet in effect, there was no difference between GLMMs that included consistency as a factor against those without, $\chi^2 = 0.82$, $p = .36$.

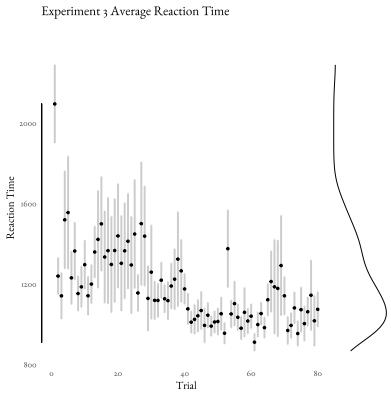

Figure 25: Reaction time during the t1 phase of Experiment 3. Black points represent mean reaction time for all participants for a trial. Lines represent standard error of the mean for each trial.

T2

Overall Accuracy. Which dimension was relevant did not affect accuracy,

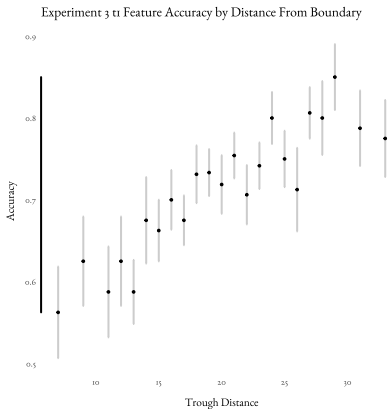

Figure 26: Experiment 3 training accuracy by relevant boundary distance.

Welch's $t(37.61) = -1.03$, $p = .3112433$, $BF_{10} = 0.16$. All further analysis will drop dimensionality as a factor.

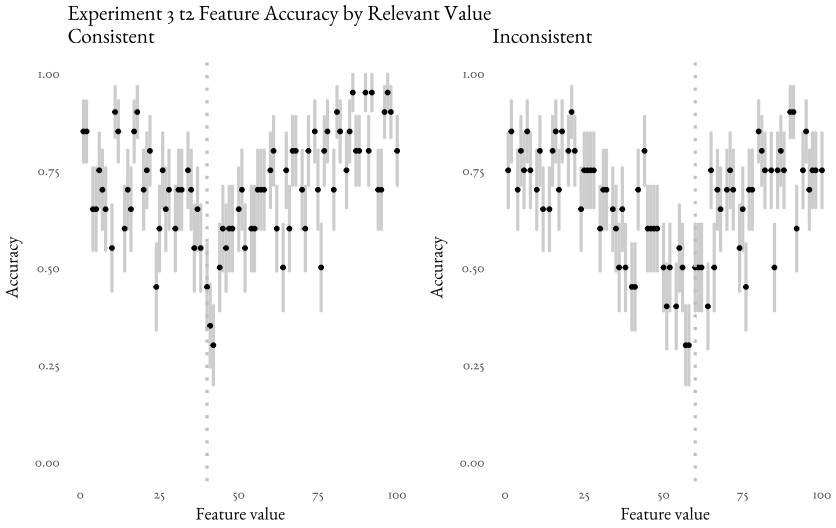

Figure 27: Experiment 3 t2 feature accuracy by feature value of the relevant dimension. The category boundaries are marked at 40 and 60.

Participants were also somewhat accurate during t2. Calculating an average accuracy score for each participant, we see $\mathcal{M}_m = 0.69$, $SE_m = 0.05$. A logistic generalized linear mixed model with accuracy as a dependent variable and consistency as an independent fixed variable and participant and trial number as random factors showed no influence of consistency on model fit compared to the null model, $\chi^2 = 1.46$, $p = .23$. A similar repeated measures ANOVA also showed no influence of consistency on accuracy, $F(1, 38) = 1.20$, $p = .28$. The same was true for a model with reaction time as the dependent variable, $\chi^2 = 0.02$, $p = .89$, $F(1, 38) = 0.02$, $p = .90$.

A GLMM examining only the first 24 items showed no effect of consistency on accuracy, $\chi^2 = 0.48$, $p = .49$. A similar model also showed consistency does not have an effect on reaction time, $\chi^2 = 0.04$, $p = .85$, $\mathcal{M}_{consistent} = 1176ms$, $\mathcal{M}_{inconsistent} = 1154ms$.

Gain score. Gain scores, measuring the difference between accuracy in t1

and t2 for each participant, showed no difference between t1 and t2 accuracy by condition, $F(1, 38) < 1$ in an ANOVA using gain score as a dependent variable and consistency as an independent variable. Overall, participants in one condition did not outperform those in another, even accounting for relative differences in accuracy. The same pattern of results arose after examining the first 24 items where consistency showed no effect on the gain score, $\chi^2 = 0.02$, $p = .90$.

Boundary Distance. A more nuanced model of participant performance should include the relative distance from the category boundary as items closer to the category boundary are harder to discriminate than stimuli with feature values further from the boundary. In a GLMM comparison, the model that included boundary distance was preferred over the random-only model, such that items nearer the relevant category boundary were more difficult than items further away, $\chi^2 = 124.1$, $p < .0001$. This same model comparison showed that consistency was also a significant contributor to model fit, $\chi^2 = 5.93$, $p = .01$, such that the consistent group had slightly higher accuracy, $M_{\text{consistent}} = .71$, than the inconsistent group $M_{\text{inconsistent}} = .67$. However, consistency and boundary distance did not interact, $\chi^2 = 0.11$, $p = .74$. A similar model comparison with reaction time as the dependent variable showed no differences by consistency, $\chi^2 = 0.03$, $p = .86$, nor any difference by relevant feature value from the category boundary, $\chi^2 = 1.06$, $p = .3$.

Examining only the first 24 items during t2 showed a different pattern of results, where only boundary distance significantly contributed to model fit, $\chi^2 = 27.89$, $p < .0001$, but neither consistency, $\chi^2 = 1.29$, $p = .26$, nor their interaction, $\chi^2 = 1.07$, $p = .30$, contributed significantly to model fit. A similar GLMM examined the effect of consistency on reaction time and found no effect of consistency, $\chi^2 = 0.17$, $p = .68$, boundary distance, $\chi^2 = 0.18$, $p = .67$, or their interaction, $\chi^2 =$

2.69, $p = .10$.

Conflict Items. In the inconsistent condition, a logistic GLMM with accuracy as the independent variable and with whether or not the feature value of a stimulus was a conflict item (between 40 and 59) or not (between 60 and 80) as a dependent variable, with the interaction between conflict and participant and trial number as random effects found that conflict contributed to model fit, $\chi^2 = 4.54$, $p = .03$. Participants were less accurate on conflict items, $M_{\text{conflict}} = .52$, than non-conflict items $M_{\text{non-conflict}} = .61$. There was no difference in reaction time, $\chi^2 = 0.17$, $p = .68$. In the first 24 items in the t2 phase, there were no differences in accuracy between conflict and non-conflict items, $\chi^2 = 0.98$, $p = .32$, nor for reaction time, $\chi^2 = 0.05$, $p = .82$. Within the consistent condition participants showed no differences in accuracy between items to the left of the consistent category boundary and items to the right of it, $\chi^2 = 2.99$, $p = .08$, nor were there any differences in reaction time, $\chi^2 = 0.27$, $p = .6$. This was also the case for the first 24 items in t2, $\chi^2 = 1.80$, $p = .18$, and $\chi^2 = 0.39$, $p = .53$, respectively.

It is also worth comparing accuracy for stimuli with relevant feature values between 40-60 in the inconsistent condition with 40-60 in the consistent condition and performing a similar GLMM with consistency as a fixed factor. There was no difference in accuracy between consistency conditions for items with feature values between 40 and 60, $\chi^2 = 1.36$, $p = .24$, nor for reaction time, $\chi^2 = 0.11$, $p = .74$. The same was also true for the first 24 items of t2.

Discussion

Experiment 2 sought to determine what people learn about an incidental feature that they then use once that feature becomes relevant in a supervised classification task. Experiment 3 sought to increase the potential for this effect to occur by

explicating certain features of the experiment to the participants, such as the number of categories, when and how the category labels change, and the presence of two distinct dimensions in the stimuli set. Experiment 3 sought to increase the chances that a participant could and would learn about the t1-incidental feature by calling their attention to the presence of two distinct dimensions and highlighting that the relevant dimension changed during the transition from t1 to t2.

The first thing to note is that the effects seen in Experiment 3 are subtle and weak. For example, in the comparison of overall accuracy between t1 and t2 contrasted by consistency condition, there is no main effect of consistency. Only when the model complexity increases to a GLMM including a continuous predictor of distance from the category boundary do we see any differences in condition. Furthermore, these effects are not present when we analyzed only the first 24 items presented during t2, though this could be due to a low number of conflict items present in the first 24 trials.

For participants in the inconsistent condition, items to the left of the inconsistent boundary (e.g., conflict items) are harder to classify than items to the right of the boundary, and participants in the inconsistent condition are less accurate overall than participants in the consistent condition when boundary distance was controlled for. To restate the Metrics section: For participants in the consistent condition, this area to the left of the category boundary should be somewhat challenging as it leads up to the category boundary, where categorization is challenging because minor differences between stimuli determine category membership, particularly around t2 stimuli located where the t1-incidental trough in the distribution was, because these small and specific stimuli ranges haven't been seen before (e.g., t2 stimuli presented where the t1 gap was between the clusters—in other words, the stimuli there are novel, but in a familiar range). For the inconsis-

tent group, however, this area represents the other side of the t1-incidental trough, but in t2 leads up to a category boundary where a mode was present in t1 at the feature value 60. It is the other side of the consistent left range block. Were nothing learned about the t1-incidental feature's *distribution*, if we compared consistent and inconsistent, accuracy on these areas to the left of their respective t2 category boundaries should be the same. However, if the people in the inconsistent condition perform worse on items with t2-relevant feature values less than the t2 category boundary, this would lend some evidence that knowledge about the distribution of feature values is interfering with performance on classification by people in the inconsistent condition.

There is at least something going on in the t1 phase that affected performance in the t2 phase; participants are not completely ignoring the t1-incidental feature as most of the literature would indicate is the case in the traditional supervised classification task. In Experiment 2 we had some evidence that t1 matters, at least in a very broad sense comparing familiar and unfamiliar stimuli. With Experiment 3 we wanted to strengthen our manipulation to see if we could find evidence of attention to some richer parts of the t1-incidental feature dimension. In Experiment 3 we seem to have succeeded in this attempt at extending Experiment 2, at least for conflict items. In particular, conflict items are more challenging for participants in the inconsistent condition than non-conflict items. This could only be the case if they had some prior knowledge or understanding of the feature distribution of the t1-incidental, t2-relevant feature distribution within this area, above and beyond merely the ability to make a discrimination between familiar and novel feature values. This finding gives some evidence to the hypothesis that people do learn about the parameters of an incidental feature during a supervised classification task.

By leading participants to attend to the two-cluster structure of the stimuli

in t1, we may have led them to ignore familiarity. Experiment 3 seems to have removed the reaction time advantage for participants in the inconsistent condition during t2. Experiment 3 differed from Experiment 2 in the instructions provided to participants before the t1 and t2 phases by explicitly pointing out the two orthogonal sets of categories, one for each phase. Inasmuch as there was a potential advantage for participants in the inconsistent condition to use a kind of novelty strategy, the effects of which would manifest as an advantage in reaction time, that advantage seems to have gone away.

There is an alternative hypothesis that should be explicated here. It is possible that participants are learning something more abstract during their experiences in the t1 phase. They saw that the relevant dimension had a cluster structure of two modes separated by a trough. They could have used this as a kind of meta-prior and applied it as an assumption about the cluster structure of the incidental feature value. Experiments 4 and 5 examine feature distributions that might rule out the use of this kind of meta-prior. Note that this is different than the complaint that what might be learned is that the feature distributions are bimodal and split in twain, such that classifications are always distributed evenly. During t2 neither the consistent or inconsistent condition have 50/50 response ratios; the consistent condition was 40% wug/60% plit, and the inconsistent condition was the opposite, 60% plit/40% wug. Another way of putting this is that a participant could learn that there are two equal-sized clusters during t1. Then during t2, even the consistent condition is not completely consistent with what is presented in t1.

Experiments 2 and 3 both presented bimodal distributions during t1 and uniform distributions during t2. Experiments 4 and 5 preserve the bimodal structure during the t2 phase to attempt to control even further any additional distributional learning that may happen during the t2 phase that could "overwrite" what

was learned during the t1 phase. Experiment 4 simply presents the same distribution during both the t1 and t2 phases and only varies the location of the category boundary during t2, an attempt at making the consistent condition more consistent.

Experiment 4

Experiments 2 and 3 attempt to probe incidental learning in t1 with a uniform distribution during t2 as a way of attenuating learning of new cluster structures during t2. However, this structure, with its uniform distribution of both uncorrelated relevant and incidental feature values, could also change any representation of what a participant learned about the t1-incidental feature distribution during t1; For participants in both conditions during Experiment 2 and 3, the stimuli and category boundaries presented in t2 were always inconsistent to some degree. Participants had to learn that the proportions of labels assigned to stimuli were not distributed evenly, as they were in t1, even in the consistent condition. Experiment 4 attempts to correct for this by presenting the exact same feature distribution during t2 as is presented during t1. This presents its own set of issues.

Experiment 4 should be considered a rough draft of Experiment 5. It is presented here for completeness. Rather than try to iron out the design asymmetries present in Experiment 2 and 3, Experiment 4 introduces asymmetry in the inconsistent condition during t2 in an attempt to root out how the distribution of t2 stimuli influence what is brought to bear from the what is learned from the feature distributions presented during t1. That is, in Experiment 4, the inconsistent boundary is truly inconsistent: both in the location of the t2 category boundary in relation to the t1-incidental trough present in its distribution, and in the

proportion of category labels to stimuli during t2 as compared with t1.

The main design feature of Experiment 4 is that rather than switch to a uniformly distributed set of feature values in the t2 phase like in Experiments 2 and 3, in Experiment 4 the stimuli remain bimodally distributed during t2. Participants in both conditions see very similar stimuli in t2 as they saw in t1 (e.g., sampled with the same parameters). However, in the inconsistent condition, the category boundary is present at the mode of the higher trough of the bimodally distributed stimuli. This presents an interesting problem for people in the inconsistent condition: their responses are heavily skewed towards one category, while the other is rarely applied. Another way of putting this is that the consistent condition is very consistent, while the inconsistent condition is very inconsistent. This potentially violates plausible priors people bring to categorization experiments, like the idea that categories mark low-frequency regions between equal-sized clusters (e.g., it is possible that Experiment 4's inconsistent condition is more difficult than the consistent condition even without a t1 phase). While this was an attempt to differentiate the inconsistent condition from the consistent condition, it has the side effect of potentially increased the difficulty of the t2 task for the inconsistent condition relative to the consistent task.

Methods

The design, materials, and procedure of Experiment 4 were very similar to those of Experiment 3, including the explicit reference to the categories dimensions. The difference between Experiment 3 and 4 derive from the difference in the feature distribution presented during the t2 phase. The t1 phase of Experiment 4 was exactly the same as the t1 phase of Experiment 3.

In Experiment 4 the t2 phase is distributed exactly the same as the t1 phase:

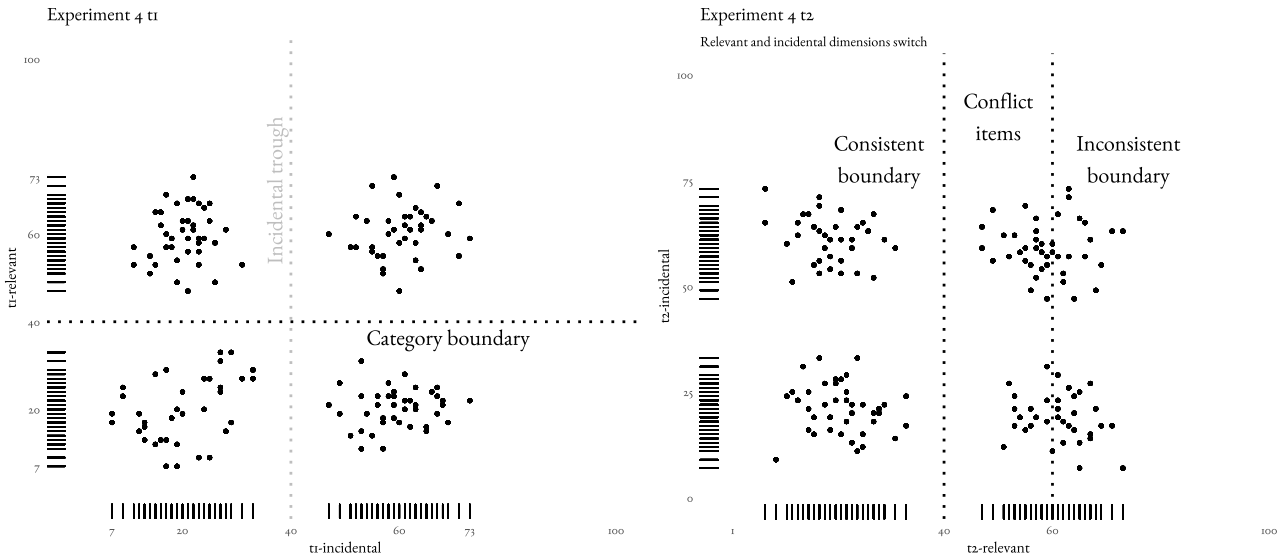

Figure 28: Experiment 4's feature distributions and category boundaries in the t1 and t2 phases.

both feature distributions are distributed bimodally with modes at 20 and 60 and a trough at 40. The category boundary for the consistent condition remains at 40; it is, in effect, a repeat of the t1 phase but with a different relevant dimension and different labels for the categories. The inconsistent condition has a category boundary at 60. As a result of the feature distribution having a mode at this value with a tail of feature values higher than that, there are much more examples of "wugs" (60) than "plits" (20) in the inconsistent condition than in the consistent.

Participants. A total of 40 participants recruited from Mechanical Turk completed the experiment. No one who participated in Experiments 1, 2, or 3 participated in Experiment 4.

Materials. The materials used were sampled from the same set of stimuli as Experiments 2 and 3. The stimuli used during t2 have the same parameters and the same lack of correlation between feature dimensions as the stimuli presented during t1. However, they are newly sampled stimuli, such that participants do not see the same relevant-incidental feature value pairs more than once in an

experiment, and those feature value pairs remain uncorrelated.

Procedure. The procedure was exactly the same as Experiment 3. Experiment 4 also used the same explicit language as Experiment 3 during the briefing sections that came before the t1 and t2 phases.

Results

T1

Ten participants performed at or below chance. Calculating an average accuracy score for each participant, we see M_{acc} = 0.72. Overall reaction time was $\mathcal{M}_{RT}$ = 1206ms. Accuracy in participants later assigned to the consistent condition was M_{acc} = .69, and accuracy in those later assigned to the inconsistent condition was M_{acc} = 75. As would be expected, because the manipulation is not yet in effect, there was no difference between GLMMs that included consistency as a factor against those without, $\chi^2 = 2.92$, $p = .09$.

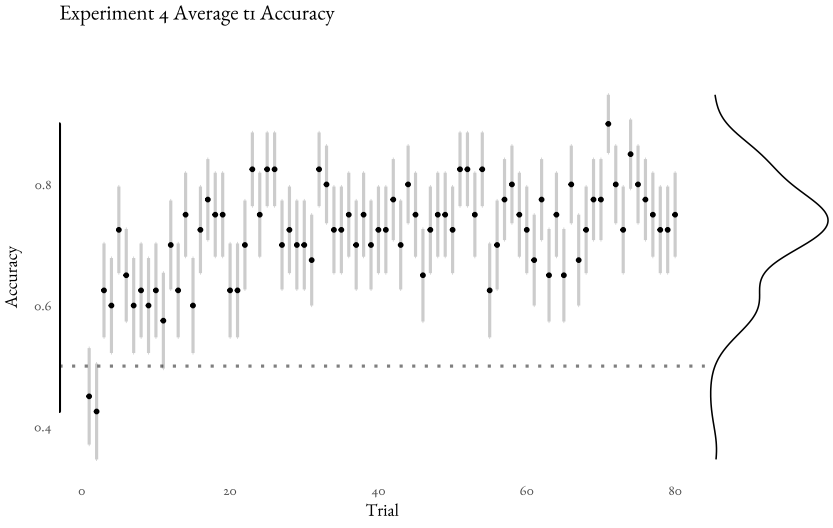

Figure 30: Accuracy during the t1 phase of Experiment 4. Black points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance accuracy for each trial.

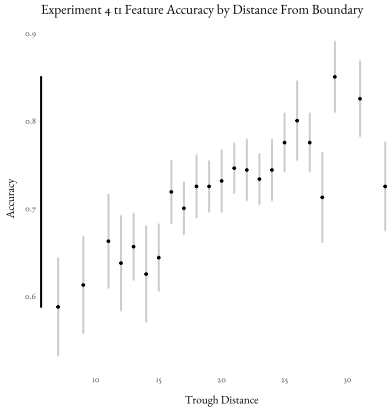

Figure 29: Experiment 4 t1 accuracy by relevant boundary distance.

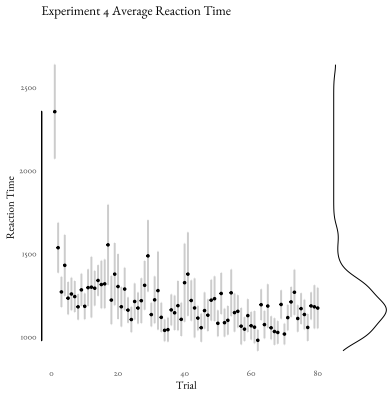

Figure 31: Reaction time during the t1 phase of Experiment 4. Black points represent mean reaction time for all participants for a trial. Lines represent standard error of the mean for each trial.

T2

Overall Accuracy. A logistic GLMM using accuracy as the dependent variable and consistency as the independent fixed factor with participant and trial number as random variables showed no better fit than the random-only model, $\chi^2 = 0.03$, $p = .87$. A similar comparison of models using reaction time as the dependent variable also showed no effect of consistency, $\chi^2 = .95$, $p = .33$. Within the first 24 items presented during t2 consistency also had no effect on accuracy, $\chi^2 = 2.83$, $p = .09$, nor on reaction time, $\chi^2 = 0.27$, $p = .61$.

Gain Score. A GLMM and ANOVA using gain score as the dependent variable and consistency as the independent fixed factor showed no effect of consistency on gain score, $\chi^2 = 1.83$, $p = .19$, $F(1, 38) = 1.78$, $p = .19$. The same was true when looking only at the first 24 items presented during t2, $\chi^2 = 1.83$, $p = .18$.

Boundary Distance. A logistic GLMM with accuracy as a dependent variable, boundary distance as an independent covariate, and consistency as an independent factor found no effect of consistency on model fit, $\chi^2 = .01$, $p = .92$, nor did the interaction between consistency and boundary distance contribute significantly to model fit, $\chi^2 = 3.04$, $p = .08$. As such, consistency had no effect on accuracy when boundary distance was accounted for. However, in the first 24 items presented during t2, consistency did have an effect on accuracy, $\chi^2 = 5.19$, $p = .02$. Boundary distance and their interaction did not contribute to model fit, however, $\chi^2 = 1.47$, $p = .23$, $\chi^2 = 1.54$, $p = .21$.

In a similar model comparison that used reaction time as the dependent variable, boundary distance did not contribute to model fit, $\chi^2 = 0.76$, $p = .38$, nor did the interaction between consistency and boundary distance, $\chi^2 = .25$, $p = .62$. But there was a main effect of consistency on reaction time, $\chi^2 = 11.85$, $p = .0006$, such that people in the consistent condition had faster reaction times than peo-

ple in the inconsistent condition, $\mathcal{M}_{\text{consistent}}$ = 1051ms and $\mathcal{M}_{\text{inconsistent}}$ = 1147ms. Within the first 24 items presented during t2, however, none of the examined independent variables contributed to model fit.

Conflict Items. Conflict items represent something different in Experiment 4. In Experiments 2 and 3, the feature distributions presented during t2 were uniform so there was ostensibly a uniform influence of t2 feature distribution on whatever t1-incidental feature representation had been built up during t1. However, in Experiment 4, because the t2 distributions are also bimodal and sampled from the same distribution, Experiment 4's conflict items should be thought of differently. One of the advantages of this measure in Experiments 2 and 3 is that the category boundaries are present for both consistent and inconsistent in the "conflict" range between 40 and 60, but in Experiment 4 there is a trough at feature value 40 for both the relevant and incidental features. This means that there are no items presented until feature value 53, which skews what this metric is actually measuring. In other words, there were fewer conflict items in Experiment 4 because the range was smaller due to the bimodal distribution of the features in t2 compared to Experiment 2 and 3's uniform distributions in t2.

In the inconsistent condition, a logistic GLMM with accuracy as the independent variable and with whether or not the feature value of a stimulus was a conflict item (between 40 and 59) or not (between 60 and 80) as a dependent variable, with the interaction between conflict and participant and trial number as random effects found that conflict did not contribute to model fit, χ^2 = 0.02, p = .89. There was no difference in reaction time, χ^2 = 0.54, p = .46. This pattern of results was borne out when examining only the first 24 items presented during t2, where conflict did not contribute to model fit where accuracy was the dependent variable, χ^2 = 0.05, p = .82, nor when the dependent variable was reaction time, χ^2 =

3.44, $p = .06$. Within the consistent condition participants' accuracy did not vary between sides of the category boundary, $\chi^2 = 0.07$, $p = .78$, nor did their reaction time, $\chi^2 = 0.01$, $p = .92$. This was also the case in the first 24 items during t2, $\chi^2 = 0.23$, $p = .63$, and $\chi^2 = 0.09$, $p = .76$, respectively.

It is also worth comparing accuracy for stimuli with relevant feature values between 40-60 in inconsistent with 40-60 in consistent and performing a similar GLMM with consistency as a fixed factor. There was a difference in accuracy between consistency conditions for items with feature values between 40 and 60, $\chi^2 = 4.27$, $p = .04$, but not for reaction time, $\chi^2 = 0.98$, $p = .32$. Furthermore, for the first 24 items presented during t2, people in the consistent condition were more accurate than people in the inconsistent condition, $\chi^2 = 4.96$, $p = .03$, but did not differ in reaction time, $\chi^2 = 0.39$, $p = .53$.

One way that the skew in the distribution of category labels in the inconsistent condition could have influenced the results is to lead particpants to be more likely to use the more common label. In the inconsistent condition, the proportion of participants' "wug" responses was .64, which is significantly above chance, $t(18) = 8.40$, $p < .001$. In the consistent condition, that proportion was .51, $t(20) = 0.56$, $p = 0.58$.

Discussion

Experiment 4 sought to maintain the bimodal t1 structure during the t2 phase as a way to increase the salience of the feature distributions of the stimuli. During the t2 phase, the category boundary for the consistent condition was located at the trough of the relevant feature distribution, but the category boundary for the inconsistent condition was located at one mode of the relevant distribution. One problem with this methodology is that the number of responses given by

participants in the inconsistent condition was heavily skewed towards the wug category because most of the stimuli presented during t2 had relevant feature values between 1 and 60.

The primary finding of Experiment 4 was that accuracy for items within the 40–60 range was higher for people in the consistent condition than people in the inconsistent condition. This finding held both overall and when we examined only the first 24 items presented during t2. Furthermore, accuracy and reaction time on conflict items was no different than non-conflict items for participants in the inconsistent condition, which is what we found in Experiment 3. On one hand, this could be taken as evidence that people learn something about the t1-incidental feature distribution and apply that to the t2-relevant feature distribution during t2. To reiterate the Metrics section: conflict items are the stimuli with relevant feature values that should be the most difficult for people in the inconsistent condition to classify because of the discrepancy between their dimension's distribution during t1 and the location of the category boundary during t2. If any intuitions are built up in the t1 phase about the incidental feature distribution, these conflict items go against those intuitions, but only in the inconsistent condition, and this is what we see in the 40 to 60 range when comparing accuracy for participants in the consistent condition to participants in the inconsistent condition. But these differences in accuracy on conflict items between the two conditions are not enough to show an effect of condition on accuracy or reaction time for all feature values, either over the course of t2 or when only examining the first 24 items presented during t2.

It is also possible, and possibly a more parsimonious explanation, that the task presented in Experiment 4 was just more difficult for participants in the inconsistent condition, but not so much so that it manifested in measures of overall

accuracy. During t1 the responses to the classification questions were split 50/50 between "feps" and "blickets," and for participants in the consistent condition this distribution continued during t2 as a 50/50 split between "wugs" and "plits" (because the consistent category boundary was at 40). But for participants in the inconsistent condition, the t2 phase presented a skewed 64/36 distribution of responses favoring "wugs." For participants in the inconsistent condition, this dissonance between the ideal t1 and t2 response ratios might have added to their already difficult task of forming a category boundary at the mode of a distribution.

We found, at least, some effect of the t1 phase on the t2 phase. Experiment 4 is somewhat informative about how people behave in a very difficult version of this dissertation's inconsistent condition: they had to contend with a skewed response ratio and a category boundary situated at one mode of a bimodal feature distribution. Experiment 5 refines the ideas behind Experiment 4 to test more explicitly and clearly whether or not participants learn the incidental feature distribution and how that affects categorization once that feature dimension becomes relevant.

Experiment 5

Experiment 5 addresses some design asymmetries present in the previous experiments. The main design goal of Experiment 5 is to present participants with either a consistent (exactly the same) or inconsistent experience (shifted 20 units) in t2. During the t1 phase, participants encounter bimodally distributed stimuli that either have a trough at 40 or a trough at 60 in both the relevant and the incidental feature distributions. Whether or not a participant is considered to be in the "consistent" condition is determined by whether or not their t2 stimuli set *also* has a trough in the same location as the t1 stimuli set. This design gets rid of any idiosyncrasies or conflations with feature values and conditions the previous experiments may harbor; gone are the differences in response proportions between t1 and t2, gone are various asymmetries between the t1 and t2 phases. Experiment 5 also addresses criticism that some results could be explained away if stimuli on one end of the feature dimensions are more or less discriminable than on the other by uncoupling the relationship between the inconsistent condition and a category boundary located at a higher feature value than the consistent condition.

Methods

Participants. A total of 40 participants recruited from Mechanical Turk completed the experiment. No one who participated in Experiments 1, 2, 3, or 4

participated in Experiment 5.

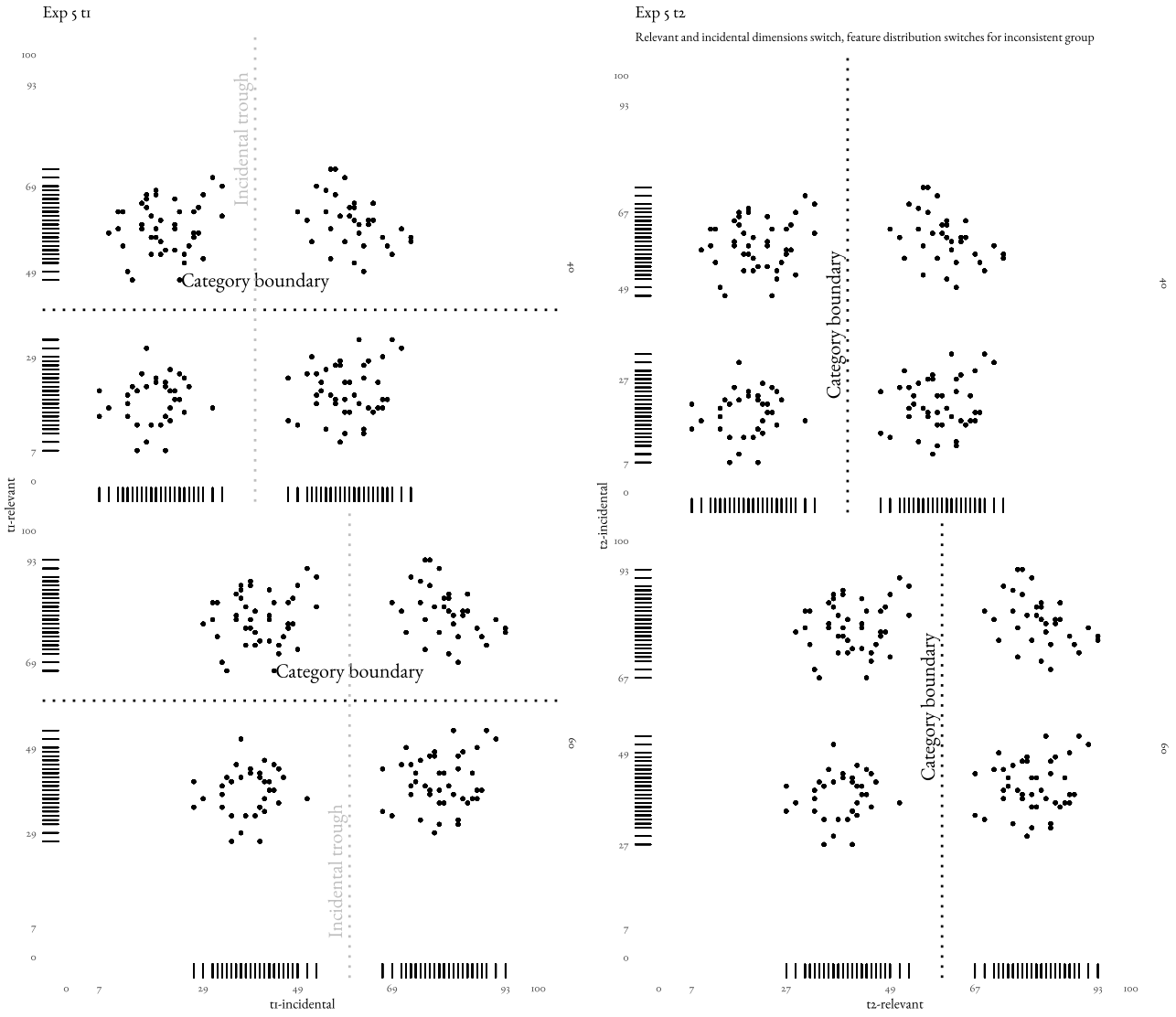

Figure 32: Experiment 5's feature distributions and category boundaries in the t1 and t2 phases. In the consistent condition participants move horizontally between quadrants from t1 to t2. In the inconsistent condition they move diagonally.

Design. Experiment 5 sought to decouple the location of the category boundary with the consistency condition. In Experiment 5, participants are assigned randomly to have a category boundary of 40 or 60 and are then independently assigned to be in either the consistent or inconsistent condition. As a result of this, Experiment 5 presents one of two bimodally distributed feature distributions for both the relevant and incidental dimensions. Both features were bimodally

distributed, like the stimuli set from Experiment 4 and the t1 sets for Experiments 2 and 3, and had feature dimensions which were uncorrelated. Each stimuli set was symmetric, too: if the category boundary for the relevant feature dimension was at 40, the inconsistent trough between the modes was also at 40. In t1 a participant saw either the stimuli set with a trough in both feature distributions at 40 or at 60. In the t2 phase, if the participant was in the consistent condition, they would see a similar stimuli set to that which they saw in t1, also with both the t2-relevant category boundary and incidental feature distribution trough at the same value. In the inconsistent condition they would see an alternative set with the feature distributions' troughs at the other value; if they had a t1-relevant category boundary and t1-incidental trough at 40, participants in the inconsistent condition would see a t2-relevant category boundary and t2-incidental trough at 60. If their t1 troughs were located at 40, then their t2 troughs were located at 60.

Materials. The stimuli were sampled from the same RFC stimuli set as Experiments 2, 3, and 4. The parameters for the feature distributions with a trough at 40 were the same as those used in Experiment 4. For those with a trough at 60, the parameters were the same, but the generated values were shifted by 20. The feature dimensions were uncorrelated like all previous experiments.

Procedure. The procedure for Experiment 5 was the same as Experiment 4. Experiment 5 also used the same explicit language as Experiment 3 during the briefing sections that came before the t1 and t2 phases.

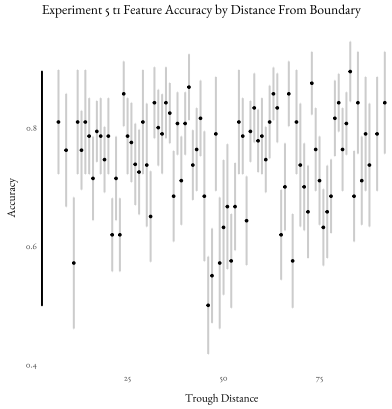

Figure 33: Experiment 5 t1 accuracy by relevant boundary distance.

Results

T1

Three participants performed at or below chance. Calculating an average

accuracy score for each participant gives M_{acc} = 0.75. Overall reaction time was $\mathcal{M}_{RT}$ = 1265ms. Accuracy in participants later assigned to the consistent condition was M_{acc} = .76, and accuracy in those later assigned to the inconsistent condition was M_{acc} = .73. As would be expected, because the manipulation is not yet in effect, there was no difference in accuracy between conditions in a GLMM that included consistency as a factor against those without, X^2 = 0.40, p = .53.

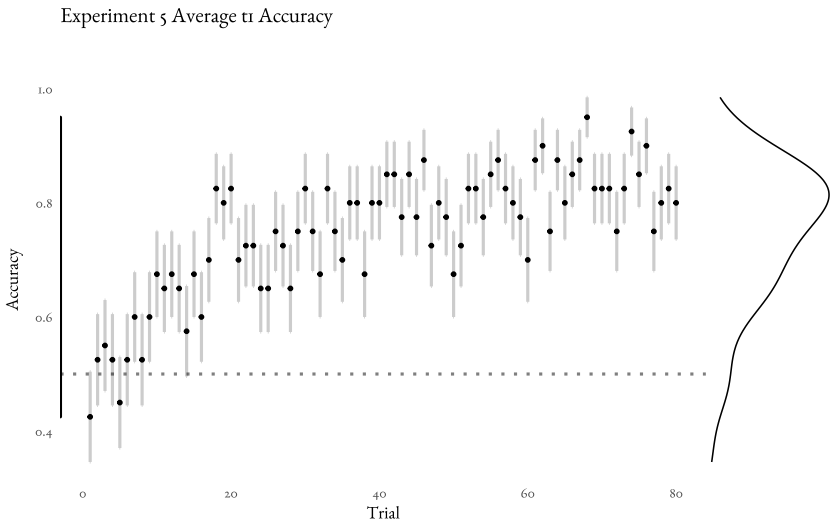

Figure 34: Accuracy during the t1 phase of Experiment 5. Black points represent mean accuracy for all participants for a trial. Lines represent standard error of the mean for each trial. The dotted gray line represents chance accuracy for each trial.

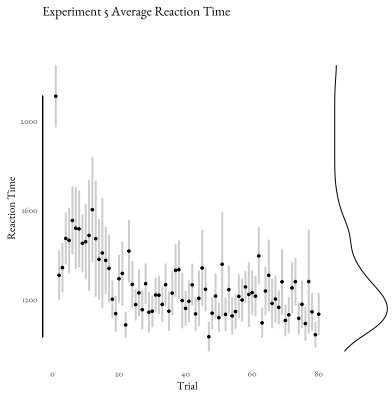

Figure 35: Reaction time during the t1 phase of Experiment 5. Black points represent mean reaction time for all participants for a trial. Lines represent standard error of the mean for each trial.

T2

Overall Accuracy. A logistic GLMM using accuracy as the dependent variable and consistency as an independent fixed factor with participant and trial number as random variables showed no effect of consistency on accuracy, χ^2 = 0.29, p = .59. However, in a similar GLMM consistency does seem to influence reaction time, χ^2 = 7.33, p = .007, such that reaction times were faster in the consistent condition, $\mathcal{M}_{consistent}$ = 982ms, than in the inconsistent condition, $\mathcal{M}_{inconsistent}$ = 1296ms. In the first 24 items presented during t2 consistency did not influence model fit when accuracy is the dependent variable, χ^2 = 1.69, p = .19, nor for reaction time, χ^2 = 1.88, p = .17.

Gain Score. A GLMM and an ANOVA with consistency as an independent factor and gain score as an dependent variable showed no effect of consistency on accuracy, $\chi^2 = 1.18$, $p = .28$, $F(1, 38) = 1.14$, $p = .29$. The same was true when examining accuracy over the first 24 items presented during t2, $\chi^2 = 1.18$, $p = .28$.

Boundary Distance. A logistic GLMM using accuracy as an independent variable, consistency as an independent factor and the relevant feature value's absolute distance from the category boundary as an independent variable found no effect of consistency on accuracy while controlling for boundary distance, $\chi^2 = 2.00$, $p = .16$, nor did the interaction between consistency and boundary distance contribute to model fit, $\chi^2 = 0.01$, $p = .93$. As in the other experiments, trough distance did increase model fit, $\chi^2 = 4.78$, $p = .03$. Interestingly, reaction time was influenced by both consistency, $\chi^2 = 85.46$, $p < .0001$, and boundary distance, $\chi^2 = 8.72$, $p = .003$, but not their interaction, $\chi^2 = 0.63$, $p = .43$.

However, in the first 24 items presented during t2, none of these independent variables influenced model fit for accuracy. But when examining reaction time on these first items, only consistency contributed to model fit, $\chi^2 = 8.50$, $p = .004$, such that participants in the consistent condition were faster to respond than participants in the inconsistent condition. This is in the opposite direction as the pattern of results for Experiment 2.

Conflict Items. Conflict items for Experiment 5 are conceptualized in a different way from the other experiments. Because a participant in the "consistent" condition saw during t2 exactly what they saw in t1, there are no novel items or even difficult ones as such (save the steady constant that items nearer to the category boundary are more difficult in almost any supervised classification task). For the inconsistent condition, however, we can frame conflict items as those items

between the t1-incidental trough and the t2-relevant category boundary. These are still items between 40 and 60; for someone in the inconsistent condition who saw a t1-incidental trough at 40 and a t2-relevant category boundary (and trough) at 60, the hardest problems should be those between the old and new category boundaries. Here the expected difficulty of conflict items is even higher: for the inconsistent condition during t2 the modes of the relevant feature are all clustered in areas where troughs were present during t1. A logistic GLMM with accuracy as a dependent variable and conflict as an independent variable found no effect of conflict on accuracy, $\chi^2 = 1.91$, $p = .17$, nor in a similar model using reaction time as the independent variable, $\chi^2 = 0.12$, $p = .73$. But on the first 24 items presented during the t2 phase, conflict did contribute to model fit, $\chi^2 = 4.52$, $p = .03$, such that accuracy on conflict items was lower than accuracy on non-conflict items, but there was no effect of conflict on a model of reaction time, $\chi^2 = 0.33$, $p = .57$. In the consistent condition, there were no differences in participants' accuracy to responses on either side of the category boundary, $\chi^2 < 0.01$, $p = .98$, nor in reaction time, $\chi^2 = 0.01$, $p = .90$. This held for the first 24 items presented during t2, $\chi^2 = 0.01$, $p = .91$, $\chi^2 = 0.37$, $p = .54$.

As in the other experiments, another measure to look at is to compare items in this 40 to 60 range within the inconsistent condition with the same items in the consistent condition. However, unlike other experiments, where the category boundary is located is now independent of condition assignment, such that half the people in the consistent condition had a category boundary of 60 and half a boundary at 40. That changes the distribution of items between participants, such that some participants in the consistent condition may not have seen the same items that people in the inconsistent condition did, and some people in the consistent condition with a category boundary at 40 may not have seen stimuli

with the same feature values that someone with a category boundary at 60 saw. A GLMM that examined accuracy within this range where consistency was an independent variable found no effect of consistency on accuracy, $\chi^2 = 0.09$, $p = .77$, but a similar GLMM found that consistency did contribute to model fit when the dependent variable was reaction time, $\chi^2 = 11.94$, $p = .0005$. Reaction times were lower for items between 40 and 60 for participants in the consistent condition, $M_{\text{consistent}} = 945$ms, than in the inconsistent condition, $M_{\text{inconsistent}} = 1375$ms. This pattern of results was borne out when examining the first 24 items presented during the t2 phase, where consistency had no effect on accuracy, but did affect model fit for reaction time, $\chi^2 = 4.23$, $p = .04$. For those first items in t2, participants in the consistent condition, $M_{\text{consistent}} = 1029$ms, responded faster than participants in the inconsistent condition, $M_{\text{inconsistent}} = 1312$ms.

Discussion

Experiment 5 represents this dissertation's most refined attempt at exploring how people learn about an incidental feature distribution and use that information in a subsequent supervised classification task. It decoupled the consistency factor with the feature's position along the distribution, such that category boundaries at 40 or 60 could both be consistent or inconsistent. It changed the consistency factor to be solely determined by whether or not the trough of the feature dimensions' distributions were in the same place as they were when presented during t1, or shifted by 20 feature values (in either direction).

The only difference in accuracy between consistency conditions was found within conflict items, where participants in the inconsistent condition were more accurate on non-conflict items, or items with feature values between 60 and 80 than on conflict items with feature values between 40 and 60, and this difference

was only present when examining the first 24 items presented during t2. This is where we predicted the effect to be, and when—early—as well.

Overall, reaction time was faster for participants in the consistent condition than for participants in the inconsistent condition, but this effect was not present when examining only the first 24 items presented during t2 and was only present when the variance introduced by boundary distance was controlled for. This difference in reaction time where participants in the consistent condition responded faster than those in the inconsistent condition was also present when the stimuli's boundary distances were accounted for, and when only examining differences in condition between conflict items, items with feature values between 40 and 60. These results add evidence to the hypothesis that people are learning something about the incidental feature distribution presented during t1, not ignoring it, and that information about that previously incidental distribution is influencing their behavior in the t2 phase.

General Discussion

When people are learning to discriminate between two categories in a supervised classification task and there is an incidental feature, they seem to encode *something* about the distribution of that incidental feature, but do not seem to accurately represent that distribution. The experiments described above seem to indicate that the distribution of an incidental feature dimension does seem to influence learning when that dimension becomes relevant. But not very much. Participants do not seem to completely ignore the incidental dimension at first, as some models of supervised classification describe, but they do not seem to fully learn or represent that incidental feature dimension in a way that maps onto its actual distribution while it is incidental. This much we knew. If there were no learning of the incidental dimension, once it became relevant the only effect we would expect to find would be one of boundary distance. Except in Experiment 2, where methodological issues probably played a role in the results, people in the consistent condition were either more accurate or faster to respond either overall, in the first 24 t2 items, or on conflict items. That is, t1 does seem to influence t2. Just *where* that consistent condition advantage was manifest was not itself consistent. In general, the consistency between the t1-incidental feature distribution and the t2-relevant distribution and classification boundary does seem to have an effect on performance. Inasmuch as performance in a classification task is indicative

of learning, it is somewhat safe to infer that participants are learning *something* about the t1-incidental feature distribution, and because of the structure of these experiments, that learning is unsupervised. This is counter to many expectations set out in the literature. Much of the category learning literature focuses less on the distributional properties of few continuous dimensions with a unimodal (one-dimensional) classification rule and more on stimuli with many binary dimensions. This study builds on literature mentioned in the introduction by designing the distribution of feature values in a such a way as to highlight the continuous nature of the dimensions. It demonstrated that it is possible for participants in a supervised classification task to learn about a task-incidental dimension.

This dissertation rests on accuracy and reaction time differences between the consistent and inconsistent conditions during t2. Experiment 1 was an attempt to replicate and validate our instantiation of the supervised classification task seen throughout the literature with stimuli developed for this dissertation. It confirmed that participants learn about the task-relevant dimension during the supervised classification task. Experiment 1 also examined what participants could learn and use from the task-incidental dimension. Participants learned about the incidental dimension though they did not seem to represent it as though they learned about its bimodal distribution. Instead, participants rated extreme values as more frequent for the relevant dimension and marginal values (nearer the center of the dimension) as more frequent along the incidental dimension. It seems like participants represented the relevant dimension as two clusters, one for each category, but represented the incidental dimension as a single cluster, rating marginal (more central) values as higher. Experiment 1 showed that participants *do* learn about the incidental dimension in a supervised classification task; they seem to overrepresent the frequency of values along the incidental dimension

that are closer to the center of the distribution, and underrepresent them along the relevant dimension. Effectively, there is a sense in which participants learned to discriminate between familiar and unfamiliar stimuli along the t1-incidental feature dimension.

Exposure to the t1-incidental feature distribution does seem to influence typicality judgments of feature values along the incidental dimension. But does it give a kind of "head start" to learning about the feature distributions once the t1-incidental feature becomes the t2-relevant feature? In Experiment 2, people in the inconsistent condition were more accurate on non-conflict items than conflict items during the first 24 trials. This could be driven by the novelty effect described below, or could be due to a violation of the representation of the t1-incidental feature distribution by the new and shifted t2-relevant category boundary. Participants in the inconsistent condition seemed to be faster at responding than participants in the consistent condition. One explanation for why this happened could be that participants in the inconsistent condition were adopting a kind of novelty strategy. From this, we infer again that participants are learning about the incidental feature during training in a way that seemed consistent with the cluster pattern seen in Experiment 1, but there was a lot of room to improve our experiment.

Experiment 3 was mostly the same as Experiment 2 with refinements to increase the chances of learning about the incidental feature value, most notably that participants were explicitly told about the experimental design and the category structures they would see during each phase of the experiment. The results showed that participants in the inconsistent condition were less accurate than participants in the consistent condition, but only when we accounted for trough distance, a finding which did not hold when we looked at only the first 24 t2 prompts. Par-

ticipants in the inconsistent condition were also less accurate on conflict items than non-conflict items, this time over all 80 trials, but not in the initial 24 trails as seen in Experiment 2. If participants learned nothing about the t1-incidental feature dimension we would expect these two sides around the t2-relevant category boundary to be equivalent. This lends some evidence that knowledge about the distribution of feature values interfered with classification performance for people in the inconsistent condition, indicating that they are learning something about the distribution of feature values along the t1-incidental, t2-relevant dimension. Experiment 3 also removed the reaction time advantage for the inconsistent condition, such that if it were being driven by a novelty strategy, that strategy has disappeared after the category numbers and dimensional change were made explicit during the experiment instructions.

In Experiment 4 the t2 phase presented the same distribution of stimuli as the t1 phase, meaning both dimensions were distributed bimodally. This was to align the distributions of t2 as much as possible with those present in t1 such that inconsistency in t2 would be foregrounded; in the previous experiments, the stimuli presented in t2 were uniformly distributed. Participants in the inconsistent condition had a category boundary located at the upper mode of the bimodal distribution. This also meant that participants in the inconsistent condition had a skewed split between the distribution of labels during t2, a confound that was addressed in Experiment 5. The main finding in Experiment 4 was that participants in the consistent condition were more accurate on items with relevant feature values between 40 and 60. This could be seen as either evidence that learning about the t1-incidental feature distribution was used during t2, *or* that the task was just more difficult for participants in the inconsistent condition because of the aforementioned skew. Experiment 5 refined the ideas of Experiment 4.

Experiment 5 decoupled feature values from the consistency condition. During t1 participants saw either the same bimodal distribution of feature values with a trough at 40 as in the previous experiments' t1 phases or one shifted with a trough at 60. During t2 a participant was in the inconsistent condition if their t2 features' distributions were shifted 20 feature values in one or the other direction and were in the consistent condition if they saw the same stimuli as they saw during t1. The main finding was that participants in the inconsistent condition were more accurate on non-conflict items (60–80) than on conflict items (40–59), and this finding only held when we examined the first 24 t2 items. Reaction time favored the consistent condition in most analyses in something of a reversal from Experiment 2. It is possible that the bimodal structure of the t2 phase gave participants something more systematic to learn during the t2 phase such that any influence of a novelty strategy was attenuated or not as salient as the uniform t2 distributions presented in Experiment 2 and 3.

Our strongest evidence of t1's effect on t2 came from conflict items. Conflict items are stimuli whose feature value and label conflict in the inconsistent condition. Within the inconsistent condition, we compared these items, which sit to the left of the inconsistent category boundary in t2, with items to the right of it. In Experiments 2 and 5 participants were less accurate on conflict items that not, but only within the first 24 items. In Experiment 3 this was also present when all the trials were examined but not in the first 24 items. Comparing items in this range between consistency conditions illustrates what was learned during t1 with which t2 interfered. For conflict items in Experiment 4, people in the consistent condition were more accurate, and in Experiment 5, people were faster to respond than people in the inconsistent condition.

The overall picture seems to be that people are learning about the incidental

feature dimension, and are learning about it or representing it in a way that is different from the relevant dimension. Where they seem to learn about the relevant dimension's distribution in a way that has something to do with its distribution from the category boundary leading to a representation of two clusters, they seem to be learning the incidental feature dimension as a single cluster centering around a central value. People do learn about the incidental feature dimension in a supervised classification task, though it is not a robust effect, and the experimental methodology (e.g., explicitly describing the number of categories and when the category criterion changes) and distribution of stimuli features affects what and how participants learn.

That is the positive picture. But one could also take the results as mostly negative. Differences in accuracy and reaction time are small when present. That could mean that most of the time in most supervised classification situations people do not really use what they learn about the t1-incidental dimension when it becomes relevant in t2, or that any dimensional representation brought about by unsupervised learning in t1 is weak and is quickly overwritten once the supervised classification task begins in t2. But the mixed bag of results, some present, some null, do not favor that explanation as parsimonious. We think it is more likely that is an artifact of our experimental designs. They are underpowered because we were iterating experimental designs to find the right paradigm.

Our consistent condition was not always that consistent. In many cases, the t1 distribution of labels was, for both conditions, different than the t2 label distribution. In every experiment, for participants in the consistent condition the t1 label distribution was 50%/50%, but in Experiments 2 and 3 it was split 60%/40%, as were the labels for the inconsistent condition. In Experiments 4 and 5 the consistent condition maintained a 50%/50% label distribution between the labeled

categories, but in Experiment 4 participants in the inconsistent condition had a very skewed distribution of labels. Experiment 5 was the only experiment where the distribution of labels between categories was evenly distributed, and where there was no change in that label distribution between t1 and t2. In that sense, we could argue that the effects of interest in Experiments 1–4 might have been attenuated by the confounding factor of skewed label distributions. But the results of Experiment 5 were not that different from those in the other experiments, so perhaps our consistent condition was consistent enough.

Additional training, or training to higher criteria, might have increased the chance for participants to learn to ignore the incidental dimension. Initially, attention to the integrated feature dimensions is spread before participants know how to solve the task and classify the stimuli. However, it could also be the case that it is only after a participant learns the t1 classification task that they then spread their attention to the incidental feature, such that incidental learning would only become solid and reliable in supervised classification after many more trials than it would take to reach near perfect accuracy on the classifying stimuli based on their relevant dimension. But that is a story for future work. Another limitation of the experiments is that they might not provide enough training for participants to learn the incidental distribution well enough. Eighty trials might not be enough exposure during t1. Because the experiments recruited from Amazon Mechanical Turk, the use of long training times might have resulted in higher participant attrition; there are few studies over 20 minutes long, and those that are that long tend to have some kind of strict admission criteria that would limit the generalizability of the results even further. In the same vein, the reward amount would necessarily need to be higher to maintain participants' attention.

One further analysis that could be done, among many, is to figure out how

many trials into t2 it takes to overwrite any effect the t1 trough distance might have had. Whatever representation participants have of the t1-incidental feature distribution, it is likely that such a representation is quickly adjusted or reformed once that dimension becomes relevant. One extension of this work would be to do that kind of analysis to determine the useful duration of t2 before running something like Experiment 5 with an extended number of stimuli presented during t1 on hundreds of people. The effect of t1 on t2 should eventually be overpowered by t2's supervised classification task such that the number of trials in t2 could be far fewer than what's required for establishing the feature distributions in t1; any bias carried over from t1 into t2 will eventually be washed out by the t2 task itself.

The experiments are underpowered for how weak the effect was. Instead of focusing on increasing our experiments' power by increasing the number of participants we decided to refine the experimental methodology by iteratively refining those methods across experiments. The mixed results of Experiment 2 motivated this decision: we realized we could potentially maximize our chance of seeing the effect of interest, if it exists, by changing the wording in the instructions. Then we experimented with changing the t2 distributions and removing as many methodological asymmetries between conditions as possible. Experiment 5 represents the last of these refinements and removed as many confounding variables introduced by the structure of the methods as possible while still maintaining the thread of general structure to access what we were interested in. There are, of course, more refinements that could be made, such as training to some criteria, or having the experiment in real time adjust the presented stimuli to maximize accuracy and learning, such as presenting items further away from the category boundary and lead up to items closer to the category boundary once a participant achieves mastery of a particular subset of the feature distribution. Ideally, we would col-

lect more data from Experiment 5, but practical constraints of funding and time intervened.

Experiment 1 highlighted the differences between separable and integrated stimuli in our tasks. Participants trained on RFCs did not favor one dimension over the other, unlike those trained on Gabor patches. Previous work showed that integrated dimensions are more likely to show task-incidental learning about within-category structure (Gureckis & Goldstone, 2008; Hoffman & Rehder, 2010), and we built on this work with more continuous stimuli with more dense feature values for a richer dimensional representation. An interesting extension to this work could involve more contrasts between integrated and separable stimuli using better, more normed separable stimuli than the basic Gabor patches used in this study. Many separable stimuli have the same issue ours did, where one dimension seems to have more salience or afford better learning than the other. Furthermore, separable stimuli often afford a kind of rule-based categorization which can be characterized in language; this is more difficult to describe in language with integrated stimuli, but can lead to more transfer between tasks (Kattner, Cox, & Green, 2016; Smith & Grossman, 2008; e.g., Smith & Sloman, 1994).

Ultimately it seems like the effects are there, but they are very small, and to see them one needs a lot of power. The distribution of the t1-incidental dimension does seem to affect accuracy or reaction time when prompting about the t2-relevant dimension, but it doesn't seem to affect it a lot. The next step is to move from this more exploratory study, which involved an iterative experimental design, to a preregistered and confirmatory design similar to that of Experiment 5, but with far more participants. This experiment would vary the number of t1 and t2 trials, too, to explore the thresholds beyond which participants ignore the

task-incidental dimension and before which they attend to it just as much as the relevant one. This single large experiment with some variance in the parameters would allow for models to be built to describe and characterize participants' representations of the feature dimensions.

The practical significance of this study is the evidence it adds to a growing body of literature that counteracts many findings in the learned inattention and blocking literature. As stated before, adults show learned inattention during extra-dimensional shifts between categories in category learning tasks (Dopson, Esber, & Pearce, 2010; Kruschke & Blair, 2000, Hoffman & Rehder, 2010). That is, they attend selectively while learning the first category and incur the cost of inattention to the previously irrelevant dimension while learning the second category if an extra-dimensional shift was present (Best & Yim, 2013). However, using continuous and rich stimuli does show evidence of learning about stimuli features that are incidental to the task (Gureckis & Goldstone, 2008; Hoffman & Rehder, 2010). This study builds on that story by using more continuous stimuli whose distribution varied in a consistent and designed way. It cannot be the case that people are simply ignoring the t1-incidental feature distribution during t1 because there were differences in accuracy and reaction time between conditions in each experiment. The simple story that persists in the literature that people ignore incidental features in the supervised classification task is incorrect, nor is it the case that people learn to ignore features incidental to the task. This dissertation explored what participants in the supervised classification task learn and how they then use what they learned in subsequent classifications.

Table of Results of Experiments 2–5

Experiment	Whole set	First 24
Experiment 2 (uniform testing)		
1. Overall Accuracy		
Accuracy	no effect	no effect of consistency
RT	inconsistent > consistent	inconsistent > consistent
2. Gain Score	no effect	no effect
3. GLMM incl. Trough Distance		
Accuracy	trough distance only	trough distance only
RT	inconsistent > consistent, nothing else	inconsistent > consistent, nothing else
4. Conflict Items		
Within inconsistent (conflict)		
Accuracy	no effect	non-conflict > conflict
RT	no effect	no effect
Between consistent and inconsistent (consistency)		
Accuracy	no effect	marginal, consistent > inconsistent, p = .08
RT	marginal, inconsistent > consistent, p = .07	marginal, inconsistent > consistent, p = .08
Experiment 3 (uniform testing)		
1. Overall Accuracy		

Experiment	Whole set	First 24
Accuracy	no effect	no effect
RT	no effect	no effect
2. Gain Score	no effect	no effect
3. GLMM incl. Trough Distance		
Accuracy	consistent > inconsistent, trough distance	trough distance only
RT	no effects	no effects
4. Conflict Items		
Within inconsistent (conflict)		
Accuracy	non-conflict (M = .61) > conflict (M = .52)	no effect
RT	no effect	no effect
Between consistent and inconsistent (consistency)		
Accuracy	no effect	no effect
RT	no effect	no effect
Experiment 4 (bimodal testing)		
1. Overall Accuracy		
Accuracy	no effect	marginal, consistent > inconsistent, p = .09
RT	no effect	no effect
2. Gain Score	no effect	no effect
3. GLMM incl. Trough Distance		
Accuracy	trough distance only	consistent > inconsistent, nothing else
RT	consistent > inconsistent, nothing else	no effects
4. Conflict Items		
Within inconsistent (conflict)		
Accuracy	no effect	no effect
RT	no effect	marginal, conflict > non-conflict, p = .06
Between consistent and inconsistent (consistency)		

Experiment	Whole set	First 24
Accuracy	consistent (M = .64) > inconsistent (M = .54)	consistent > inconsistent
RT	no effect	no effect
Experiment 5 (bimodal testing)		
1. Overall Accuracy		
Accuracy	no effect	no effect
RT	consistent > inconsistent	no effect
2. Gain Score	no effect	no effect
3. GLMM incl. Trough Distance		
Accuracy	trough distance only	no effect
RT	consistent > inconsistent, interaction w/ trough distance	consistent > inconsistent, nothing else
4. Conflict Items		
Within inconsistent (conflict)		
Accuracy	no effect	non-conflict > conflict
RT	no effect	no effect
Between consistent and inconsistent (consistency)		
Accuracy	no effect	no effect
RT	consistent > inconsistent	consistent > inconsistent

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, *55*(1), 11–27.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. Com's mechanical turk. *Political Analysis*, *20*(3), 351–368.

Best, C. A., & Yim, H. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119.

Bott, L., Hoffman, A. B., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General*, *136*(4), 685–699.

Brooks, L. R., Squire-Graydon, R., & Wood, T. J. (2007). Diversion of attention in everyday concept learning: Identification in the service of use. *Memory & Cognition*, *35*(1), 1–14.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data. *Perspectives on Psychological Science*, *6*(1), 3–5.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software, Articles*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216–226.

Drucker, D. M., & Aguirre, G. K. (2009). Different spatial scales of shape similarity representation in lateral and ventral loc. *Cerebral Cortex (New York, NY)*, *19*(10), 2269–2280. doi:10.1093/cercor/bhn244

Gureckis, T. M., & Goldstone, R. L. (2008). *The effect of the internal structure of categories on perception proceedings of the 30th annual conference of the cognitive science society* (pp. 1876–1881). Cognitive Science Society Austin, TX.

Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: Towards a unified developmental approach. *Developmental Science*, *13*(4),

622–635.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340.

Ipeirotis, P. (2010). *Demographics of mechanical turk [working article]*. New York University.

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.

Kattner, F., Cox, C. R., & Green, C. S. (2016). Transfer in rule-based category learning depends on the training task. *PLoS ONE*, *11*(10), e0165260.

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

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science Connection Science*, *8*(2), 225–248.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083–1119.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830–845.

Leeuw, J. R. (2014). JsPsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Levering, K. (2012). *Generative processing as a framework for human category learning* (PhD thesis). State University of New York at Binghamton, Ann Arbor.

Levering, K., & Kurtz, K. (2014). Observation versus classification in supervised category learning. *Mem Cognit*.

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among us- and india-based workers on mechanical turk. *Behav Res*, *47*(2), 519–528.

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.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613.

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

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems 14* (Vol. Advances in Neural Information Processing Systems 14, pp. 841–849). MIT Press.

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

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci*, *4*(12), 1244–1252. Retrieved from http://dx.doi.org/10.1038/nn767

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

Schwartz, D. L., Bransford, J. D., Sears, D., & others. (2005). Efficiency and innovation in transfer. *Transfer of Learning from a Modern Multidisciplinary Perspective*, 1–51.

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

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

Smith, E. E., & Sloman, S. A. (1994). Similarity-based versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386.

Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, *10*(4), 502–532.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432. doi:10.1007/s11222-016-9696-4

Ward, T. B. (1980). Separable and integral responding by children and adults to the dimensions of length and density. *Child Development*, *51*(3), 676–684.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 908–923.

Xu, J.-M., Zhu, X., & Rogers, T. T. (2012). Metric learning for estimating psychological similarities. *ACM Trans. Intell. Syst. Technol.*, *3*, 55:1–55:22.

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