

Effects of Labelling on Object Perception and Categorisation in Infants

Arthur Capelier-Mourguy, MSc

PhD Thesis

Supervisors: Gert Westermann, Katherine E. Twomey

November 2019

Abstract

How do labels impact object perception and enhance categorisation? This question has been the focus of substantial theoretical debate, particularly in the developmental literature, with conflicting results. Specifically, whether labels for objects act as additional perceptual features or instead as referential pointers to category concepts has been the subject of intense debate. In this thesis, we attempted to shed a new light on this question, combining empirical results on both infants and adults, and neurocomputational models.

First, we developed a dual-memory neurocomputational model of long-term learning inspired by Westermann and Mareschal's (2014) model, to test predictions of the two main theories on labelling and categorisation on existing infant data, and to generate predictions for a follow-up study. Our modelling work suggested that for the empirical designs considered and age groups tested, labels were processed as object features, as opposed to having a more referential role.

We then focused on explicitly testing potential attentional effects of auditory labels during categorisation in an empirical study. More precisely, we studied the interaction between feature salience, feature diagnosticity, and auditory labels, in a categorisation task. Surprisingly, we found that 15-month-old infants and adults could learn labelled categories in which the salient feature (head of line-drawn novel animals) was non-diagnostic of category membership, but the non-salient feature (tail) was, without adopting a different pattern of looking compared to participants in a control group. Although our data did not provide clear evidence for a true null effect, this finding was once again more compatible with the theory that labels act as features, not referents. This finding also led us to reconsider the use of eye movements and looking times as a proxy for learning, as it seemed that participants could learn more without looking more.

Given our empirical results on salience and diagnosticity of features, and given the methodological differences in the handling of feature salience and diagnosticity in the categorisation literature, we developed a simple auto-encoder model to further study the impact of salience differences between features in the context of a categorisation task, with or without a label. Our simulations suggested that bigger disparities in salience between different features of an object can result in differences in terms of learning speed and compactness of categories in internal representations, hinting that future empirical studies should consider feature salience in their design.

Overall then, this thesis provides some evidence in favour of the *labels-as-features* theory through the use of empirical eye-tracking data on infants and adults, and neurocomputational modelling. This thesis further asks new questions on the importance of feature salience in categorisation tasks, and the interpretation of eye movement and looking time data in general.

Contents

| | |
|---|------------|
| Acknowledgements | vi |
| Declaration of Contributions | vii |
| Paper 1: Modelling the Effect of Learned Labels in Infants | vii |
| Paper 2: Unmatched Feature Salience and Diagnosticity | vii |
| Paper 3: Labels Drive Adults' Attention to Salient Features | viii |
| Paper 4: A Model of Labelling and Attentional Focus | viii |
| Agreement and Declaration | viii |
| 1 Literature Review | 1 |
| 1.1 Labelling and Categorisation in Infants | 1 |
| 1.1.1 <i>Labels-as-Symbols</i> | 3 |
| 1.1.2 <i>Labels-as-Features</i> | 5 |
| 1.2 Labelling and Categorisation in Adults | 7 |
| 1.3 Computational Models of Categorisation | 8 |
| 1.3.1 An Overview of Different Modelling Approaches | 9 |
| 1.3.2 Westermann and Mareschal's Dual-Memory Model (2012, 2014) . . | 12 |
| 2 Modelling the Effect of Learned Labels in Infants | 15 |
| 2.1 Introduction | 16 |
| 2.2 Experiment 1 | 18 |
| 2.2.1 Model Architecture | 18 |
| 2.2.2 Procedure | 21 |
| 2.2.3 Results | 22 |
| 2.2.4 Discussion | 24 |
| 2.3 Experiment 2 | 25 |
| 2.3.1 Stimuli | 25 |
| 2.3.2 Procedure | 26 |
| 2.3.3 Results | 26 |
| 2.3.4 Discussion | 28 |
| 2.4 General Discussion | 30 |
| 3 Unmatched Feature Salience and Diagnosticity | 35 |
| 3.1 Introduction | 36 |

| | | |
|----------|---|------------|
| 3.2 | Experiment 1 | 39 |
| 3.2.1 | Methods | 39 |
| 3.2.2 | Results | 43 |
| 3.2.3 | Discussion | 50 |
| 3.3 | Experiment 2 | 52 |
| 3.3.1 | Methods | 52 |
| 3.3.2 | Results | 53 |
| 3.3.3 | Discussion | 58 |
| 3.4 | General Discussion | 59 |
| 4 | Labels Drive Adults' Attention to Salient Features | 64 |
| 4.1 | Introduction | 65 |
| 4.2 | Methods | 69 |
| 4.2.1 | Data Handling and Software Specifications | 69 |
| 4.2.2 | Participants | 69 |
| 4.2.3 | Materials | 70 |
| 4.2.4 | Procedure and Design | 71 |
| 4.3 | Results | 72 |
| 4.3.1 | Behavioural Results | 73 |
| 4.3.2 | Eye-tracking Results | 75 |
| 4.4 | Discussion | 83 |
| 5 | A Model of Labelling and Attentional Focus | 89 |
| 5.1 | Introduction | 90 |
| 5.2 | Methods | 93 |
| 5.2.1 | Model Architecture | 93 |
| 5.2.2 | Stimulus Encoding | 94 |
| 5.2.3 | Procedure | 95 |
| 5.2.4 | Data Handling and Software Specifications | 96 |
| 5.3 | Results | 97 |
| 5.3.1 | Familiarisation | 97 |
| 5.3.2 | Contrast Test Trials | 101 |
| 5.4 | Discussion | 102 |
| 6 | General Discussion | 109 |
| | Bibliography | 116 |

Acknowledgements

The work presented in this thesis was funded by a Leverhulme Trust Doctoral Scholarship, and with the support of the Lancaster University Psychology Department and Lancaster Babylab. I would like to thank all the parents that let me test their lovely babies, and all the adult participants who took part in my research.

I would also like to thank Prof. Gert Westermann and Dr. Katherine E. Twomey for their dedication and patience supporting me through these four years of PhD studies. Thanks in particular to Gert for giving me the opportunity to work on this project and pushing me when I needed it, and to Katie for voluntarily giving me so much of her time when she had no obligation to. I would also like to thank Dr. Katharina Kaduk for creating such a great work environment at the Babylab, and all of my colleagues for making my time in Lancaster so much better.

I would particularly like to thank Dr. Katherine E. Twomey for introducing me to R and new statistical tools, Dr. Robert Davies for further supporting my training in those statistical tools, and countless people online for broadening even further my knowledge and understanding of statistics (on StackOverflow, CrossValidated, GitHub).

Finally, thanks to everyone else who supported me throughout this long journey: my friends in the Folk Society, and those at the Irish session in town, for all the good music; Benjamin and Arnaud in their far off lands for the amazing trips together and the crazy chats; Sébastien for all the time we spent together in conferences and visiting each other; Vicky for her devoted support when I needed it the most; Jess for giving me new impetus to reach the finishing line; and Beth for giving me something to look forward to after my PhD. And of course my family, who have always been there for me, and in particular my mum for her unconditional love and support. I could not have done it without you all.

Declaration of Contributions

Paper 1: Modelling the Effect of Learned Labels in Infants

Study concept *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Coding *Arthur Capelier-Mourguy*

Statistical analysis *Arthur Capelier-Mourguy*

Results interpretation *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Manuscript drafting *Arthur Capelier-Mourguy*

Manuscript revisions *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Contribution of principal author: 85%

Paper 2: Unmatched Feature Salience and Diagnosticity

Study concept *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Study design *Arthur Capelier-Mourguy*

Data collection *Arthur Capelier-Mourguy*

Statistical analysis *Arthur Capelier-Mourguy*

Results interpretation *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Manuscript drafting *Arthur Capelier-Mourguy*

Manuscript revisions *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Contribution of principal author: 85%

Paper 3: Labels Drive Adults' Attention to Salient Features

Study concept *Arthur Capelier-Mourguy*, Gert Westermann, Ho Yeung¹

Study design Ho Yeung, *Arthur Capelier-Mourguy*

Data collection Ho Yeung²

Statistical analysis *Arthur Capelier-Mourguy*

Results interpretation *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Manuscript drafting *Arthur Capelier-Mourguy*

Manuscript revisions *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Contribution of principal author: 85%

Paper 4: A Model of Labelling and Attentional Focus

Study concept *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Coding *Arthur Capelier-Mourguy*

Statistical analysis *Arthur Capelier-Mourguy*

Results interpretation *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Manuscript drafting *Arthur Capelier-Mourguy*

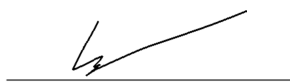
Manuscript revisions *Arthur Capelier-Mourguy*, Gert Westermann, Katherine E. Twomey

Contribution of principal author: 85%

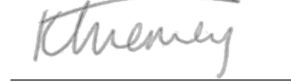
Agreement and Declaration

We the undersigned agree with the above stated proportions of work undertaken for each of the above manuscripts contributing to this thesis. I the main author declare that this thesis is my own work and has not been submitted in substantially the same form for the award of a higher degree elsewhere.

Arthur Capelier-Mourguy



Katherine E. Twomey



Gert Westermann



¹third year undergraduate student co-supervised by Arthur Capelier-Mourguy and Gert Westermann

²data collected as part of the student's research project, but reanalysed and interpreted in more depth as part of this thesis

Chapter 1

Literature Review

To begin this thesis, we review the existing empirical and computational modelling literature on categorisation and specifically on the effects of auditory labels on categorisation. First, we present the different theories for the effect of categorisation early in development, the main focus of this thesis. We then present the converging evidence for the role of category labels in adults. Finally, we briefly describe a few different modelling approaches for human labelled categorisation, before extensively presenting the neurocomputational model around which we will build our subsequent modelling work.

1.1 Labelling and Categorisation in Infants

From a very young age, infants learn to group objects into categories. They do so by considering the features that are similar between objects within the same category only and less similar to the features of other out of category items (Mareschal & French, 2000; Mareschal et al., 2000). To assess categorisation in pre-linguistic infants, a novelty preference looking time procedure is usually conducted after a training phase (familiarisation or habituation): looking times between a within-category and an out-of-category items presented together are recorded. The preference to either the familiar (within-category) or novel (out-of-category) item according to different conditions has been well documented (see Houston-Price & Nakai, 2004; Oakes, 2010, for a review). In short, the more infants have encoded an item/category, the more likely they will be to look at a more novel out of category item at test than a member of the previously familiarised category. In habituation studies, in which the item/category is presented until a criterion of lack of interest is reached as indicated by significantly shorter looking times to the screen, infants are then expected to look more at the novel stimulus at test. On familiarisation tasks, in which the item/category is presented for a fixed number of times, the meaning of a novelty or familiarity preference is less clear. In such paradigms, if the familiarisation lasted long enough for infants to fully encode the stimuli, infants should exhibit a novelty preference. Conversely, if they did not have enough time to fully encode the stimuli during familiarisation, then infants should exhibit a familiarity preference for the stimulus they have already partially encoded. As such, results in familiarisation studies should be analysed at different time frames during familiarisation to allow a more

fine-grained analysis of the preference results, for example using the reduction in total looking time between the beginning and end of the training as a measure of habituation. It is also good practice to correct for natural preference for presented items, instead of looking at preference against chance alone (e.g. Oakes, 2010; Quinn, 2004).

Other paradigms can be used to test for categorisation learning in infants; although they generally rely on the same intrinsic mechanisms and measures, they all shed a different light on the processes at hand. Looking time studies are typically limited by the available dimensions for the presentation of stimuli, and do not account for the multimodal aspect of categorisation learning. Addressing this issue, paradigms exist with 3D objects in place of 2D pictures, allowing for a richer manipulation and encoding by the infant during the familiarisation phase, and in particular allowing researcher to test for the importance of haptic information such as object texture during categorisation (e.g. Graham & Diesendruck, 2010). For such paradigms, infants behaviour are recorded on video and later coded by hand. A first straightforward way of using physical objects is to replicate the structure of the visual familiarisation task described above: first infants are familiarised with exemplars from different categories one at a time, then are presented at test with a new exemplar from an old category and an entirely new stimulus (e.g. Mandler & McDonough, 1993; Oakes et al., 1997; Oakes et al., 1991; Younger & Furrer, 2003). Looking times are here replaced by handling and looking behaviours. Another alternative offered by the use of physical objects is the possibility to present infants with a set of objects belonging to two different categories, rather than presenting them with one exemplar at a time (e.g. Rakison & Butterworth, 1998). This paradigm is particularly interesting as it allows for more active comparisons between both within-category and out-of-category exemplars. In this paradigm, touching sequence and length of each manipulation can both be recorded and used as indices of infants' cognitive processes. Typically, categorisation is evidenced by successive touching of within-category exemplars, though earlier stages may involve more alternating between the two categories than within-category successions (Oakes & Plumert, 2002), much in the same way that both familiarity and novelty preference can both be seen in visual preference studies depending on context. Finally, the use of physical objects allows for functionality-based categorisation, where the experimenter shows a particular action with an object to an infant, then presents the infant with an object of the same category and a different category, and records whether or not the infant generalises the action to the correct object (e.g. Mandler & McDonough, 1996; Träuble & Pauen, 2007). In the rest of this thesis, we will focus on visual preference methods.

In addition to these studies of visual category learning, auditory features have also been shown to impact categorisation. A first effect that has been documented is the overshadowing of processing of visual features by auditory stimuli: when presented with a visual and a non-linguistic auditory stimulus with simultaneous onset, 4-year-old children preferentially attend to the auditory stimulus, reducing the amount of resources devoted to the visual stimulus, thus leading to a poorer encoding of this visual stimulus than in a silent presentation (Sloutsky & Napolitano, 2003). This effect arises from the

earlier maturation of the auditory compared to the visual sensory system, given that the auditory system starts functioning during the last trimester of gestation (Birnholz & Benacerraf, 1983; Jusczyk, 2000). Auditory overshadowing has been replicated and extended to auditory labels in 8- and 12-month-old infants; precisely, although auditory stimuli enhanced attention as evidenced by total looking times, infants were more likely to form categories in silence (Robinson & Sloutsky, 2007a). Further, this auditory overshadowing effect, as measured by differences in visual processing speed, disappeared in 14-month-old infants when using familiar sounds, namely, human speech, non-linguistic sounds embedded in human speech in place of words, and non-linguistic sounds on which infants were pre-familiarised (Robinson & Sloutsky, 2007b). Finally, it has been argued that, initially, this auditory overshadowing might help categorisation by reducing the level to which visual features of objects are represented, thus reducing the dissimilarities between exemplars (Robinson & Sloutsky, 2004; Sloutsky & Robinson, 2008). Further, in addition to this overshadowing effect, there is evidence that sounds in general enhance attention overall, as measured by longer total looking times to stimuli when these stimuli are presented together with an auditory stimulus, compared to presented in silence (e.g. Roberts & Jacob, 1991; Robinson & Sloutsky, 2007a). This might also help infants encode objects and learn categories.

However, there is evidence that linguistic auditory inputs are more effective at helping categorisation, even in pre-linguistic infants, compared to other auditory inputs, even though auditory inputs in general enhanced infants' attention (Balaban & Waxman, 1997). This facilitatory effect of human speech sounds has been extended to communicative sounds in general, such as content-filtered speech sounds (Balaban & Waxman, 1997), onomatopoeic sounds (Roy, 2003), or even chimpanzee vocalisations (Ferry et al., 2013). Further, there is evidence that “meaningful” environmental sounds help categorisation, for example the sound of a dog barking would help forming a category for dogs (Hendrickson et al., 2015). Additionally, category exemplars are often encountered in real life with their corresponding name, and such labelling events have been shown to specifically improve categorisation (e.g. Althaus & Westermann, 2016; S. A. Gelman & Coley, 1991; Gliga et al., 2010; Graham & Poulin-Dubois, 1999; Plunkett et al., 2008). Despite numerous studies on the question of the mechanisms by which such auditory labels help categorisation in infants, no converging evidence has been found, and two main theories still attempt to tackle this question: the *labels-as-symbols* theory, and the *labels-as-features* theory (for an overview of this debate, see S. A. Gelman & Waxman, 2009; Sloutsky, 2009; Waxman & Gelman, 2009).

1.1.1 *Labels-as-Symbols*

On the *labels-as-symbols* account, label representations are from an early stage of development qualitatively different from object representations, with labels acting as privileged referential markers for categories in a top-down way (Waxman & Markow, 1995). According to this theory, labels help infants to form categories by highlighting the *diagnostic features* of these categories, that is, features that are shared by within-category

exemplars but not shared by out-of-category exemplars. For example, knowing that both llamas and rabbits have four legs and are fluffy is not helpful to discriminate them into two categories, while the long neck of llamas and the big ears of rabbits are both diagnostic features for their respective categories. In the study that gave rise to this theory, Waxman and Markow (1995) found that 12- to 13-month-old infants could reliably form basic-level categories (e.g. cows vs. dinosaurs) without labels. They could however only form superordinate-level categories (e.g. animals vs. vehicles) when provided with a label. Following this first study, it has been shown that the addition of a label allowed 10-month-old infants to form categories they would not otherwise form, either to group together into one category a set of items they would otherwise divide into two categories (Plunkett et al., 2008), or alternatively to divide into two categories a set of similar exemplars that were accompanied by two different labels (Althaus & Westermann, 2016). A further two studies argued that infants grouped a set of dinosaurs into one super-ordinate category only when hearing a linguistic label, but not when hearing tones that reproduced the rhythm of the labelling phrases (Ferry et al., 2010; Fulkerson & Waxman, 2007). However, these studies did not include a silent control condition, and a subsequent replication attempt showed that infants formed the same category in silence as they did when hearing a label, highlighting the importance of control conditions to evidence a true effect (Chen & Westermann, 2012). Thus, the original results are best explained as being due to tones blocking, but not labels allowing learning of the category.

Recently, two studies used eye-tracking on 12-month-old infants to study the online process of categorisation and how labelling impacted it, addressing the question of *how*, not just *how well*, auditory labels affect infants' categorisation (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015a). In both studies, infants were familiarised with one set of two-featured stimuli forming one category, either in silence or paired with a single label. Importantly, both features were made equally salient for all stimuli, and one feature's shape varied more than the other's. Those studies then demonstrated that the presence of a label during familiarisation induced a focus on low-variability features early during familiarisation (Althaus & Mareschal, 2014), and increased and sustained attention to familiarised within-category versus novel out-of-category low-variability features in a subsequent test phase in silence (Althaus & Plunkett, 2015a). The label here drove infants' attention towards the low-variability features, and changed how the category was encoded in memory, with more importance being given to those low-variability features in infants' internal representations. These studies however assumed that infants did form a category representation without explicitly testing for it by using two contrasting categories and assessing successful learning of the category labels at test. They further assumed that low-variability features represented diagnostic features, drawing on the idea that diagnostic features are common to all members of a category, thus of relatively low variability within the category. Thus, these studies leave open the question of how labels would impact object representations when learning multiple categories with clearly diagnostic features. Indeed, while diagnostic features might be of relatively low

variability, low-variability features are not always diagnostic. For example, the tail that dogs, cats, chimpanzees, and many other animals all share is of very low variability, but is not diagnostic for any of those categories.

1.1.2 *Labels-as-Features*

Conversely, the *labels-as-features* theory assumes that label representations are integrated into object representations (Sloutsky & Fisher, 2004). On this account, labels have no special status, but contribute to object representations in the same way as other features such as shape and colour: a dog is an animal with four legs, a tail, fur, a muzzle, that barks, and is called ‘dog’. According to this theory, labels help categorisation in infants by adding to the overall similarity between category exemplars, being a highly reliable diagnostic feature. Crucially, this theory predicts that, with a similarity-based categorisation mechanism, infants would need to consider all features to compute the similarity between two items. Thus, this theory expects infants, in a categorisation task, to remember individual exemplars and their particular features. On the opposite, with a knowledge-based categorisation mechanism in which labels act as symbolic markers highlighting diagnostic features, infants would only need to focus on those diagnostic feature and could ignore the other, non-informative features. Further, infants would not need to compute a precise similarity measure, and as such, could rely on a general, prototypical representation for the diagnostic features. As a result, infants’ ability to recall or recognise individual exemplars and their particular features should be drastically reduced. Sloutsky and Fisher (2004) first confirmed that infants were able to recall particular category exemplars and their features, as predicted by their *labels-as-features* theory. They further reproduced their empirical findings with a simple mathematical model of inter-exemplar similarity treating the label as a feature amongst others, although weighing more in the comparison process—in other words, the label was more *salient*.

In a more recent paper supporting this theory (Deng & Sloutsky, 2012), 4- to 5-year-old children were first familiarised with five-featured anthropomorphic stick figures divided into two labelled categories. Importantly, the salient head was diagnostic of category membership for all exemplars, and was animated to further increase its salience. They then had to complete two tasks: infer a category label for a new exemplar (categorisation task), or infer a missing feature for a new labelled exemplar (induction task). Crucially, in the induction task, some exemplars were given the label and most features corresponding to one category (*A*), but exhibited a salient feature (the head) corresponding to the opposite category (*B*). For these exemplars with a conflicting label and head, the authors predicted that, if infants saw labels as symbolic markers, they would infer category *A*, but if they saw labels as features, they would likely infer category *B* in accordance with the head, a feature that was more salient and equally diagnostic compared to the label. Their predictions were upheld, with infants consistently inferring missing features as corresponding to the same category as the head.

Using the same categorisation/induction paradigm, another study showed that not

only identical but merely similar auditory labels contributed to the judgement of within-category similarity of 5-year-old children (Sloutsky & Fisher, 2012). More precisely, the study considered the rate at which items that were more or less similar to a set of familiar or trained category exemplars, were labelled as belonging to that same category. They then compared this rate of category label inferences to the rate at which labels that were more or less similar to a familiar or trained category label induced feature inferences corresponding to that same category. They noted that those two rates were similar, that is, between-exemplar similarity in terms of visual features or auditory labels had the same impact on categorisation mechanisms, suggesting that labels are not different from other object features.

While the *labels-as-features* theory does not make any assumptions of specific attentional effects of labels, an important prediction of the *labels-as-symbols* theory is that auditory labels will drive attention towards diagnostic features. Directly testing for this effect, a study first familiarised 6- to 8-month-old infants with two labelled categories, but halfway through the experiment changed the to-be-learned category to another category where the previously diagnostic features were no longer relevant (Best et al., 2013). If labels do direct attention towards diagnostic features, then we would expect infants to exhibit a switch cost when the features they were previously focusing on are no longer diagnostic. This was not the case however, suggesting that labels do not direct attention towards diagnostic features as predicted by the *labels-as-symbols* theory.

Taking a closer look into attentional processes during category learning, was an eye-tracking study on 8- to 12-month-old infants using the same five-featured category prototypes as in Deng and Sloutsky (2012), and building two categories from those prototypes by changing one feature at a time, including the salient head (Deng & Sloutsky, 2015). More precisely, they compared the effect on attention of hearing an auditory label (label condition) and seeing the feet moving (motion condition), when learning one of the two categories, the other only being used for contrast at test. Their main finding was that infants learned the category only in the motion condition, not in the label condition, a result conflicting with previous studies on categorisation in infancy, which all showed a positive effect of labelling on category learning. This facilitatory effect of having a dynamic visual feature was explained by an increase in distributed attention, as seen by an increase in the number of visual shifts between different features. The authors further claimed that labels failed to attract attention to commonalities. However, no one feature was more diagnostic for category membership than any other in their stimuli, and as such, there was no one feature that the label could have highlighted across exemplars. Conversely, infants did exhibit longer looking to the head in the label condition, suggesting that the label did drive attention to a feature that is arguably highly diagnostic in real-life categories, a result that could thus be interpreted as evidence in favour of the *labels-as-symbols* theory.

Overall then, studies addressing the question of the role of auditory labels in categorisation early in development have been myriad, using a variety of paradigms. This

field is however still understudied, and possible explanations of the conflicting evidence observed might lie in yet unknown effects. For example, it was suggested that the timing in the presentation of an auditory label and visual stimulus was more important than earlier thought (Althaus & Plunkett, 2015b). In this study asynchronous presentation of the visual stimulus followed by the label led to a positive effect of labelling on categorisation abilities, whereas synchronous presentation of both stimuli led to an auditory overshadowing like effect and the absence of a facilitatory effect of labelling on categorisation.

Another such effect that has not been accounted for and thus has not been controlled for consistently across studies is the salience of different features, with only one study looking at salience maps of familiar stimuli and how 4- and 12-month-old infants' looking patterns compared to those salience maps (Althaus & Mareschal, 2012). However, this study only revealed that throughout familiarisation infants looked less at the salient features and more at other features. The authors explain this as attention being driven by bottom-up processes at first, governed by the salience of different features, to become more top-down controlled with an active information-seeking behaviour. However, this result could be equally explained in terms of habituation to the better-encoded high-salience features at the beginning, leading to a novelty preference looking at other less salient features later on. Nonetheless, this study gave evidence that feature salience has an impact on feature preference and encoding, and on categorisation. This result calls for further studies addressing the question of the effect of auditory labels on categorisation with feature salience as one of the controlled parameters of the design.

1.2 Labelling and Categorisation in Adults

If the question of the role of labels in infancy is still debated, it is generally agreed upon that labels act as symbolic markers in adults. In an early study, Sloutsky et al. (2001) noted that 4- to 5-year-old children responded reliably according to the *labels-as-features* theory, 11- to 12-year-old children's responses were more consistent with the *labels-as-symbols* theory, and 7- to 8-year-old children were seemingly in a transitional phase, with some participants responding in a feature-oriented fashion, while others were more relying on the label. Further, most studies providing evidence for the *labels-as-features* theory in infants contrasted these findings with evidence that adult controls treated labels as symbolic category markers. For example, in the studies we mentioned on infants that tested adults at the same time, adults exclusively relied on identical labels to define categories (Sloutsky & Fisher, 2012), showed a switching cost when previously diagnostic features became irrelevant (Best et al., 2013), and most adults made more label-consistent feature inferences (Deng & Sloutsky, 2012). However, in this last study some adults showed either head-consistent inferences or a mix of the two behaviours. Studies have also been conducted specifically on adults to understand the mechanisms by which labels act as symbolic markers to help categorisation.

First, it is clear that even redundant labels help adults to learn categories more efficiently, both in terms of reaction times for category membership decisions, and in terms

of quicker increase in accuracy compared to a control condition with no auditory label (Lupyan et al., 2007). In a subsequent study, redundant labels were shown to increase the detection speed of exemplars belonging to the corresponding category (Lupyan & Spivey, 2010), hinting at a priming effect of auditory labels on category concepts.

One study looked more precisely into the priming effects of auditory labels and other meaningful auditory cues such as the barking of a dog or the word “barking” (Lupyan & Thompson-Schill, 2012). They showed that hearing a label activated category representations in a more effective way and more consistently between subjects than did other auditory cues, linguistic or not. This work was further extended to show that participants, when hearing non-linguistic auditory cues, activated the concept of a specific category exemplar in a specific sound-producing action, whereas participants activated more general, decontextualised, category representations when hearing a label (Edmiston & Lupyan, 2015). These results echo early evidence that category labels (e.g. “dog”), but not labels relating to object features (e.g. “snout”), are treated differently from other object features, visual or not (Yamauchi et al., 2007), and had a stronger effect on the induction of a missing feature (Yamauchi & Yu, 2008).

Finally, tapping directly into cerebral processes, a study showed that inducing an enhancement of the cerebral mechanisms linked with labelling improved the formation of “sparse” categories, that is, categories defined by only a few diagnostic features (Perry & Lupyan, 2016). Specifically, they enhanced labelling mechanisms by up-regulating activity over Wernicke’s area, involved in language comprehension, via transcranial direct current stimulation. Conversely, disturbing labelling mechanisms by down-regulating activity over Wernicke’s area improved the formation of more multi-dimensional categories with no fully diagnostic features (Perry & Lupyan, 2014).

1.3 Computational Models of Categorisation

Computational modelling is an essential tool in cognitive sciences, allowing us to implement theories and assumptions these theories make, test them in a controlled environment, and understand what aspects of the theories tested impact predictions in which way. There are two philosophies when it comes to modelling: building complex models of human cognition that account for a variety of task results, or building simpler models that account only for a certain type of task. While complex models might seem more appealing, their very complexity makes it hard to identify and understand the underlying mechanisms at play. For example, a recent deep neural network model replicated the emergence of a shape bias in categorising information (Ritter et al., 2017). Precisely, this result evidenced that a powerful regularity extractor, when extensively trained to group real world stimuli into labelled categories, learned to give more importance to the shape rather than colour of objects when building new categories in a subsequent test phase. That is, there is a shape bias in the way humans structure real-life objects into labelled categories. Crucially, this model does not explain if such a shape bias emerged for one reason or another, and thus only replicates its existence without explaining it. Conversely, simpler models might seem more limited, for example by the nature and size

of their input, often not as ecological. Nonetheless, they allow us to understand what aspects of the model lead to the observed results, and how changes in model parameters might relate to different results and different aspects of the theory they are built on.

1.3.1 An Overview of Different Modelling Approaches

Many models have been used to study categorisation and the effects of labels on categorisation, ranging from self-organising maps (SOM; Mayor & Plunkett, 2010), simple similarity-based mathematical models (SINC; Sloutsky & Fisher, 2004), clustering algorithms (SUSTAIN; Love et al., 2004), connectionist models with objective encoding of rules and features (ATRIUM/ALCOVE; Erickson & Kruschke, 1998; Kruschke, 1992), and many others. They all give insights into different processes of categorisation, at different levels of abstraction.

Sloutsky and Fisher (2004) proposed SINC, for “Similarity, INduction, and Categorization”, a model that considered a simple view of categorisation, based on similarity. In this model, objects were considered in terms of a fixed number of features (shape of the head, eyes, ears, etc.), with a finite set of possible values for each feature. To compare two objects, their similarity was computed based on the number of features matching between them. Each feature could further have a different weight, meaning that mismatches on different features would have more or less of an impact on the computed similarity value. In this model, labels were treated like other features, weighing more in the similarity decision. Although very simple and abstract, this model succeeded in replicating a broad range of empirical studies. It thus suggests that categorisation is to some extent a function of feature-by-feature similarity between encountered objects, with labels acting on the same level as other features.

A more realistic model, combining a self-organising map with Hebbian learning, two biologically plausible mechanisms, accounted for the developmental shift in the role of labels, from treating them as features to seeing them as symbolic markers for categories (Mayor & Plunkett, 2010). SOMs are used to encode complex, often multi-dimensional, stimuli into a two-dimensional grid of neurons, with each neuron representing a particular exemplar from the input space, and neighbouring neurons coding for similar exemplars once the SOM is fully trained (Kohonen, 1990). Thus, the presentation of an input to a SOM will activate a cluster of neurons depending on how similar their receptive field is to the new input. Given two pre-trained SOMs with linguistic labels and objects typically encountered by infants, this model could first generalise the link between a label and the corresponding category after a single label-object presentation, and, over time reinforce the link between label instances and object instances, so that labels slowly became predictors of objects rather than merely associated with them. These associations were learned via a Hebbian learning over the connections between the two SOMs, reinforcing a connection when an object was presented together with a label to the model. Since the Hebbian learning happened as a function of neuron activation, and since each input (linguistic or visual) activated a cluster of neurons at different levels, this model slowly learned to generalise those label-object connections into label-category connections. Al-

though this model was limited by the size of the two SOMs, and the fact that they both required background training before being connected, it provided evidence that a simple model, based on biologically plausible components that are unimodal SOMs and cross-modal Hebbian learning, could account for the developmental trajectory of the role of labels in categorisation.

Crucially, these models of categorisation did not implement any attention mechanisms, one of the key components thought to underlie infants' and adults' categorisation behaviours. The first model of categorisation to do so was ALCOVE (Attention Learning COVERing map, Kruschke, 1992), combining an exemplar representation with perceptron-inspired error-driven back-propagation learning over three layers. In this model, stimuli were divided in multiple dimensions (height, colour, etc.), each of which could have different values, and were coded by a separate input unit each whose activation was the corresponding feature's value. This input layer then propagated to a hidden layer, in which each unit represented a previously encountered exemplar, with a receptive field on each dimension in multidimensional psychological space. As such, the model was initialised with no hidden units, and those were added one at a time as the model encountered new exemplars. Then, those hidden units were activated by the new stimulus depending on their similarity to this new stimulus; more precisely, their receptive field over each input dimension responded with exponential decay. Crucially, an attentional gating parameter shaped the width of all those receptive fields for each dimension, allowing the model to learn to give more or less importance to specific features, effectively distorting its representation of the world. Finally, those hidden exemplar units connected to a categorical decision output layer, making ALCOVE a supervised model of category learning. All the parameters were then updated by backpropagation of the error. Although this model initially used only previously encountered exemplars as hidden units, it could be initialised with a full covering grid of hidden units, simulating long-term background knowledge.

In ATRIUM, a later model, ALCOVE was combined with a parallel rule-based categorisation model (Erickson & Kruschke, 1998). The two parallel networks competed to make a categorisation decision, and thus learned for each input if a rule-based or an exemplar-based approach was better suited for this type of input. These two models (ALCOVE and ATRIUM) accounted for a great many categorisation task results. Particularly, ALCOVE replicated tasks that previous models had failed to replicate because of their lack of an attention mechanism, for example when successful categorisation depended on correlated dimensions of the input stimuli (Medin et al., 1982). ATRIUM further replicated tasks in which some categorisation decisions were rule-based and others exemplar-based, as is the case for example for the 'mammal' category in which most exemplars can be categorised based on their similarity with other mammals, but whales and dolphins call for a rule-based categorisation. However, a different set of parameters was necessary for each result reproduction, reducing the model's explanatory and predictive power.

Another model inspired by ALCOVE, SUSTAIN (Supervised and Unsupervised STrat-

ified Adaptive Incremental Network), was however able to address this shortcoming, reproducing all the same data with a single set of parameters (Love et al., 2004). The key difference between ALCOVE and SUSTAIN was that SUSTAIN replaced hidden exemplar units by ‘cluster neurons’. Those neurons were structurally similar to the exemplar units we described earlier, with a receptive field over each dimension of the multidimensional stimulus space. However, when an exemplar unit was added for each new stimulus in ALCOVE, in SUSTAIN, cluster neurons were simply updated with every encountered stimulus, to better fit the data. This model still learned incrementally, creating a new cluster when encountering a new stimulus that could not be accounted for by the model, and centring this new cluster on this new stimulus. Interestingly, this model could do so either in a supervised or unsupervised way. When supervised, a new cluster was created when a queried dimension (or a category label) was falsely predicted. When unsupervised, a new cluster was simply added when the closest cluster to the new stimulus was not close enough, or put differently, when the neuron representing this cluster was not activated above a pre-determined threshold. SUSTAIN successfully replicated empirical data from adults on supervised classification learning but also on a broader range of tasks and conceptual functions linked to categorisation: learning categories at different levels of abstraction, inferring a missing object feature when hearing a label, and unsupervised category learning. Although SUSTAIN achieved a remarkable fit to human data, it remained limited by its explicit coding of stimuli: the input data were not raw, but cut into set features that could take only set values. Even though it is safe to assume that humans visual processing is capable of extracting abstract features and representations from visual inputs, this however meant that SUSTAIN was not autonomous, and was limited by experimenter bias on the coding of explicit stimulus dimensions.

One type of model that partly addresses the problem of experimenter coding bias are auto-encoders. Those models reproduce input patterns on their output layer by comparing input and output activation after presentation of training stimuli and computing the error between these two representations, then using this error to adjust the weights between units using back-propagation (Rumelhart et al., 1986). One important aspect of auto-encoders is that their hidden layers are of reduced size compared to their input/output layers; thus, auto-encoders learn to compress information in the most effective, lossless way. In doing so, they essentially extract features from complex stimuli, and are therefore well suited for categorisation tasks. Furthermore, the error-driven learning of those models matches the idea that infants learn, when presented with a novel stimulus, by comparing it to an internal representation of the same stimulus (e.g. Charlesworth, 1969; Cohen, 1973). The bigger the discrepancy between representations, the more need for information processing, and thus for a longer looking time, referred to as a *novelty preference* (see Oakes, 2010; Quinn, 2004). Thereby, such neurocomputational models have successfully captured looking time data from infant categorisation tasks (Mareschal & French, 2000; Westermann & Mareschal, 2004), using error on the network’s output layer as a proxy for infant looking times.

A recent model of categorisation in infancy was built as a dual-memory three-layered

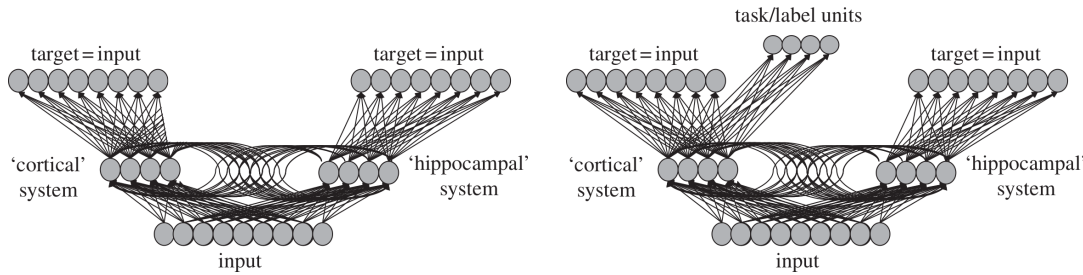


Figure 1: Structure of the Dual-Memory Network model without label output units on the left (Westermann & Mareschal, 2012), and with label output units on the right (Westermann & Mareschal, 2014).

auto-encoder neural network (Westermann & Mareschal, 2012, see Fig. 1). We discuss this model further below.

1.3.2 Westermann and Mareschal’s Dual-Memory Model (2012, 2014)

This model consisted of two simple auto-encoders with three layers (input, hidden, and output), coupled by and interacting through their hidden units. The two auto-encoders had different learning rates, and implemented on an abstract level a short-term memory (STM) and a long-term memory (LTM). The LTM component used a learning rate of 0,001 so that it encoded information relatively slowly; the STM used a learning rate of 0,1 and encoded information relatively quickly. The two auto-encoders further interacted through their hidden layers: those hidden layers were updated in parallel, receiving activation from their input layer and from one another, until both hidden layers had converged to a stable state (i.e. the change in their activation between two steps of the loop fell under a pre-determined threshold). Activation was then propagated to each output layer, and the difference between these outputs and the input (the network’s prediction error) was used to update each network’s connection weights via backpropagation of this error. Specifically, the horizontal connection from the LTM hidden units to the STM hidden units were updated using the STM’s learning rate, and vice versa, so that the influence of each component over the other was learned at the same rate as the rest of the impacted network. This model was trained over a wide range of natural stimuli, to emulate to some extent infants’ background knowledge: 19 exemplars from 19 basic-level categories taken from 4 superordinate categories (furniture, animals, vehicles, and humans), encoded through 19 meaningful features (based on object geometry and characteristics). This model was first used to replicate empirical data on the effect of background knowledge on pre-linguistic categorisation in young infants (Bornstein & Mash, 2010), in which infants familiarised faster to new exemplars of a category in the lab only if they had been habituated with different exemplars of this category at home in a two-months-long background training phase.

This model was later extended to account for the effect of labels on categorisation processes and its change over the course of development (Westermann & Mareschal, 2014, see Fig. 1). The training set for this extended model consisted of 208 exemplars

from 26 natural categories, falling into the same four subordinate categories as previously, and encoded in the same way. This model represented labels as additional output units on the LTM component only. This represented the empirical finding that infants activate learned long-term label representations when encountering category exemplars (Mani & Plunkett, 2010). Since the model did not have an input label to compare to its prediction over those units, these units were used in a supervised way, and the model had to learn to predict the correct category label depending on its input. During background training, those units were used only half of the time, accounting for the fact that, in real life, objects are not reliably labelled in every instance on which infants encounter them. Crucially, this model shed new light on the debate over the role of labels on categorisation. Here, labels were not treated as other features but were nonetheless embedded into object representations with those other features through the process of backpropagation. Labels did not have an abstract attention-driving role either. Nonetheless, when trained with labels, the model grouped categories into more compact clusters, as represented in its hidden layer, than when trained in silence. In other words, adding a label increased the perceived similarity between exemplars within a category relative to between-category similarity, as predicted by the *labels-as-features* theory, without treating the labels as other features. Thus, this model offered a new *compound-representations* account to explain early and later labelled categorisation and its developmental course.

In conclusion, we have seen that there is an ongoing debate on the role of verbal labels for categorisation in infancy. On the one hand, the *labels-as-symbols* theory argues that labels can actively guide categorisation by highlighting diagnostic features from an early developmental stage. Conversely, the *labels-as-features* theory argues that labels are first perceived as object features with no distinct role, and are simply a highly salient feature that adds to the similarity between exemplars within a category. Finally, both theories agree that labels have a more symbolic value in adults, acting as category markers, and the *compound-representations* theory offers an account of this developmental switch, supported by a neurocomputational model.

One key question that we raised in our literature review is that of the role of feature salience in categorisation, and its possible interaction with auditory labelling. We will attempt to answer this question later in this thesis, using empirical work on pre-linguistic infants and adults, and neurocomputational modelling methods. Specifically, we will study how categories in which a salient feature is non-diagnostic but non-salient features are diagnostic are learned, and how adding an auditory label changes the way these category are learned.

First, however, we extended the dual memory model with linguistic units described above (Westermann & Mareschal, 2014) to implement the assumptions of the *labels-as-features* theory. We validated this implementation by replicating existing empirical data on infants, particularly, we teased apart two theories that were equally able to explain the data: the initially implemented *compound-representations* theory, and the

labels-as-features theory that we newly implemented. We further used this model to make predictions for an ongoing follow-up empirical work. This first computational work serves as a stepping stone to our thesis, allowing us to test the explanatory and predictive power of this neurocomputational model architecture.

Chapter 2

Modelling the Effect of Learned Labels in Infants

The following chapter has been published as

Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2018). Neuro-computational models capture the effect of learned labels on infants' object and category representations. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1. <https://doi.org/10.1109/TCDS.2018.2882920>

The full text is reproduced here, however the formatting has been adapted to fit better within the thesis. This includes the use of the APA style instead of the IEEE numeric style, the rescaling of figures, and the re-writing of Table 1 into a narrower format.

This chapter was further edited after the *Viva Voce* by request of the examiners. These corrections were however minor, and only included (a) the deletion of an inexact citation, (b) changing the word “neuron” to “unit” to avoid any confusion.

Neurocomputational models capture the effect of learned labels on infants' object and category representations

Arthur Capelier-Mourguy, Katherine E. Twomey, and Gert Westermann
Lancaster University, UK

Abstract

The effect of labels on non-linguistic representations is the focus of substantial theoretical debate in the developmental literature. A recent empirical study demonstrated that ten-month-old infants respond differently to objects for which they know a label relative to unlabeled objects. One account of these results is that infants' label representations are incorporated into their object representations, such that when the object is seen without its label, a novelty response is elicited. These data are compatible with two recent theories of integrated label-object representations, one of which assumes labels are features of object representations, and one which assumes labels are represented separately, but become closely associated across learning. Here, we implement both of these accounts in an auto-encoder neurocomputational model. Simulation data support an account in which labels are features of objects, with the same representational status as the objects' visual and haptic characteristics. Then, we use our model to make predictions about the effect of labels on infants' broader category representations. Overall, we show that the generally accepted link between internal representations and looking times may be more complex than previously thought.

Keywords: connectionist model, representational development, label status, language development, cognitive development

1 Introduction

The nature of the relationship between labels and non-linguistic representations has been the focus of recent theoretical debate in the developmental literature. On the *labels-as-symbols* account (Waxman & Gelman, 2009; Waxman & Markow, 1995), labels are symbolic, conceptual markers acting as privileged, top-down indicators of category membership, and label representations are qualitatively different to object representations. In contrast, the *labels-as-features* view assumes that labels have no special status; rather, they contribute to object representations in the same way as other features such as shape and color. More recently, Westermann and Mareschal (Westermann & Mareschal, 2014) suggested a *compound-representations* account in which labels are encoded in the same representational space as objects and drive learning over time, but do not function at the same level as other perceptual features. Rather, they become closely integrated with object representations over learning and result in mental representations for objects that reflect both perceptual similarity and whether two objects

share the same label or have different labels. This approach therefore takes a middle ground between the *labels-as-symbols* and the *labels-as-features* views in that labels do not act at the same level as other object features (acknowledging that language is special as in *labels-as-symbols*), but that an integrated object representation is formed through the association between perceptual object features and labels (as in *labels-as-features*). However, despite substantial empirical work (e.g. Althaus & Mareschal, 2014; Althaus & Plunkett, 2015; Gelman & Coley, 1991; Gliga et al., 2010; Sloutsky & Fisher, 2004, 2012; Twomey & Westermann, 2017b; Westermann & Mareschal, 2014)) and a handful of computational investigations (e.g. Gliozi et al., 2009; Mirolli & Parisi, 2005; Westermann & Mareschal, 2014)), there is no current consensus as to the status of labels in object representations, and the debate goes on.

A variety of studies have demonstrated that language does affect object encoding and representations early in development. When and how in development this relationship emerges is less clear. For example, labels can guide online category formation in infants and young children (Althaus & Westermann, 2016; Graham & Poulin-Dubois, 1999; Plunkett et al., 2008), and previously learned category representations affect infants' online visual exploration in the lab (Bornstein & Mash, 2010; Hurley & Oakes, 2015), but until recently the link between learned labels and category representations had not been directly tested. Gliga et al. (2010) recently explored electroencephalogram (EEG) neural responses to stimuli in 12-month-old infants presented with a previously labeled object, a previously unlabeled object, and a new object. They found significantly stronger gamma-band activity only in response to the previously labeled object, and this, in line with previous EEG work, was interpreted as a marker of stronger encoding of this object. Twomey and Westermann (2017b) extended this work by training 10-month-old infants with a label-object mapping over the course of one week. Specifically, parents trained infants with two objects during three-minute play sessions, once a day for seven days, using a label for one of the objects, but not for the other. After the training phase, infants participated in a looking time task in which they were shown images of each object in silence. Testing the hypothesis that (previously learned) labels would affect infants' object representations, the authors predicted that infants should exhibit different looking times to the labeled and unlabeled objects. Their predictions were upheld: results showed a main effect of labeling, such that infants looked longer at the previously labeled than the unlabeled object (see Fig. 1 for the original data).

These data shed light on the debate on the status of labels. Specifically, they support both the labels-as-features and the compound-representations theories. On the labels-as-features account, if a label is an integral part of an object's representation, when the label is absent there will be a mismatch between that representation and what the infant sees in-the-moment (equally, a similar response would be expected when another of the object's features, for example color, differed from the learned representation). Since infants are known to engage preferentially with novel stimuli (Fantz, 1964; Houston-Price & Nakai, 2004), this mismatch will elicit a novelty response, indexed by increased looking times to the previously labeled object. On the compound-representations view,

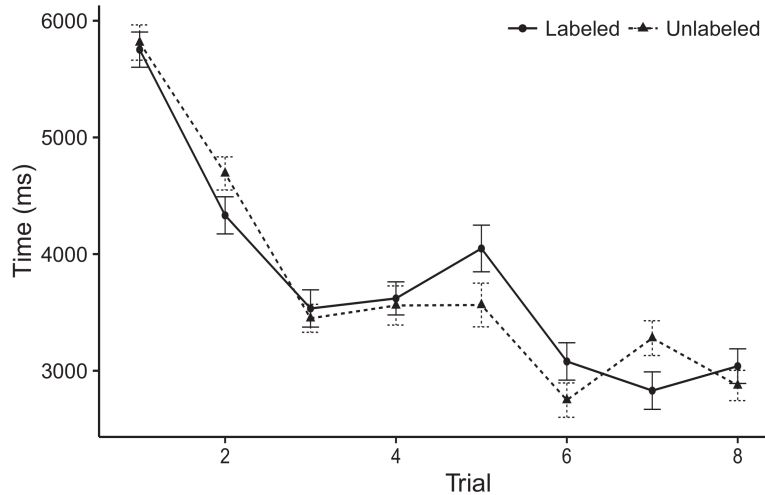


Figure 1: Looking time results from Twomey and Westermann (2017b). Error bars represent 95% confidence intervals.

seeing the previously labeled object would activate the label representation (Mani & Plunkett, 2010). This active label representation would, in turn, lead to a priming-like increase in looking time towards the previously labeled object (Baldwin & Markman, 1989; Mani et al., 2012; Mani & Plunkett, 2011).

Importantly, while the behavioral data presented in Twomey and Westermann (2017b) support either of these views, they cannot differentiate between the two. Computational models, on the other hand, allow researchers to explicitly test the mechanisms specified by these theories against empirical data. Specifically, simple computational models, by stripping back mechanisms to a minimum, allow us to precisely understand these mechanisms and discover which ones are relevant and which ones are not (for similar arguments, see McClelland, 2009; Morse & Cangelosi, 2017). Thus, here we implemented both accounts in simple computational models to explore which of the labels-as-features and compound-representations accounts best explains Twomey and Westermann’s (2017b) looking time data.

2 Experiment 1

2.1 Model Architecture

We used a dual-memory three-layer auto-encoder model inspired by Westermann and Mareschal (2014) to implement both the labels-as-features and the compound-representations theories. Such neurocomputational models have successfully captured looking time data from infant categorization tasks (Mareschal & French, 2000; Twomey & Westermann, 2017a; Westermann & Mareschal, 2004, 2012, 2014). Auto-encoders reproduce input patterns on their output layer by comparing input and output activation after presentation of training stimuli, then using this error to adjust the weights between units using back-propagation (Rumelhart et al., 1986).

Our model consisted of two auto-encoders coupled by, and interacting through, their

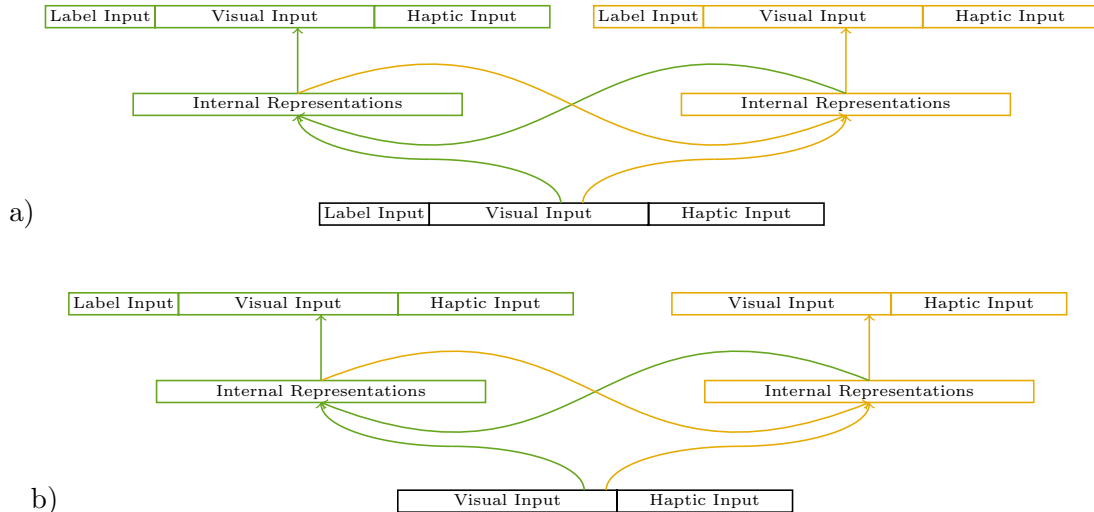


Figure 2: Structure of the Dual-Memory Network models: the Long-Term Memory is in green (left), and the Short-Term Memory in yellow (right). Layer width corresponds to number of units: 5 label, 10 visual, 8 haptic, and 15 hidden units. a) Labels-as-Features model. b) Compound-Representations model.

hidden units. These two subsystems represented, on an abstract level, a short-term (STM) and a long-term (LTM) memory component. This model has previously been used to simulate the impact of infants' background category knowledge acquired in everyday life (represented in long-term memory) on lab-based looking time experiments involving in-the-moment knowledge acquired in familiarization-novelty-preference studies (represented in short-term memory) (Westermann & Mareschal, 2014). It was therefore well suited to simulate the effects of infants' learning about objects and labels at home on their subsequent looking behavior in the lab as in Twomey and Westermann (2017b).

The two auto-encoders had different learning rates: the LTM component used a learning rate of 0.001 so that it encoded information relatively slowly; the STM used a learning rate of 0.1 and encoded information relatively quickly. For the interaction between the two networks' hidden units, both hidden layers were updated in parallel, receiving activation from their input layer and the other network's hidden layer until both hidden layers had converged to a stable representational state, with the lateral interaction resulting in no further update in their activation. The weights from the STM to LTM were treated as part of the LTM network and updated with a learning rate of 0.001; similarly, the weights from the LTM to the STM were treated as part of the STM network and updated with a learning rate of 0.1. Thus, the influence of the other memory on each network was updated at the same rate as the rest of the network. Both networks received identical input. The details for all the model parameters and the full code are available on-line (<https://github.com/respatte/LabelTime>).

2.1.1 Labels-As-Features Model (LaF)

Fig. 2a depicts the LaF model. To represent the label as a feature that was equivalent to all other features, we included it both at the input and the output level for both

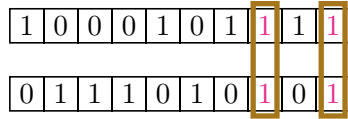


Figure 3: Encoding of stimuli, with overlapping units highlighted.

components. Thus, the label had exactly the same status as all other features in the model’s representation.

2.1.2 Compound-Representations Model (CR)

Fig. 2b depicts the CR model. Here, labels are represented only on the output side of the LTM network. Thus, in effect, the model learns to associate the perceptual object description with the label. This approach reflects the empirical finding that presenting an object to infants activates their (learned, long-term) representation of the label for that object (Mani & Plunkett, 2010).

2.1.3 Stimuli

Our stimuli were encoded as sets of abstract binary features that were designed to reflect the visual, haptic and label characteristics of the 3D object stimuli used in Twomey and Westermann (2017b). Thus, our encoding can be interpreted as a list of dummy variables that could generalize to alternative stimuli, coding for the presence/absence of one particular dimension of the stimuli (e.g. “is made of wood”, “is red”, would be plausible dimensions for the stimuli considered here).

Visual Input Twomey and Westermann’s (2017b) empirical study stimuli were two small wooden toys: a castanet, and two wooden balls joined with a string. One toy was painted red and the other blue, with color counterbalanced across children. Thus, the stimuli were visually dissimilar, but both consisted of two wooden components connected with string/elastic. To reflect the partial overlap in visual appearance of these objects, we encoded the visual component of our stimuli as patterns of activation over 10 units; each object had the same number of active units (6), with two out of the ten units active for both objects to represent commonalities between stimuli (see Fig.3).

Haptic Input As well as visual experience, infants in Twomey and Westermann (2017b) received haptic input when handling or mouthing the stimuli. We reasoned that the degree of overlap in this input would vary between infants. Because both objects were wooden and presented simultaneously, infants would have experienced some overlap in haptic experience with the objects. On the other hand, because the objects had different affordances, this overlap would never have been total. Thus, we encoded haptic input over eight units, with overlap varying randomly between two and six units between simulations. Haptic stimuli were presented to the model simultaneously with the visual stimuli and encoded in an identical fashion.

Label Input Label input consisted of five binary units, activated (set to 1) for the labeled object only. For the unlabeled object, the units were simply set to 0.

2.2 Procedure

In line with the experimental study in Twomey and Westermann (2017b), our procedure consisted of two phases. First, to simulate the 3D object play sessions at home, we trained the models with both objects, one with a label and one without a label (*background training*). Then, we simulated the second, lab-based part of the study by familiarizing the models with both objects without the labels to simulate the silent familiarization phase of the empirical study. Specifically, we ran each architecture in a familiarization phase in which the label units were inactive for both stimuli: the label inputs for the LaF architecture were set to zero, and the label outputs were ignored for both architectures (therefore not contributing to network error nor impacting on further weight updates).

To collect an amount of data consistent with infant studies, we ran a total of 40 model subjects for each architecture.

2.2.1 Play Sessions

To reflect the likely differences in playing time across children, the total number of iterations for which the model received each stimulus during background training was selected randomly from a normal distribution of mean 2000 and standard deviation 200. Stimuli were presented individually in alternating fashion. Although this does not precisely reflect the rich, combined play with both objects for different times experienced by infants, alternating the stimuli allows the model to learn more efficiently from a purely computational point of view, and should not influence results, as different training orders for the same stimuli asymptotically converge to the same solution.

2.2.2 Familiarization Training

Before familiarization training, we added noise to the STM’s hidden-to-output weights (by adding a value in the range $\pm[0.1, 0.3]$ to the existing weight values) to simulate the likely memory decay from infants’ final play session, which had taken place the previous day. Then, the label input units were set to zero, and the output units ignored, not taking them into account when computing network error and back-propagation. Haptic input and output units were also set to zero, to reflect the absence of haptic experiences in the lab experiment.

Familiarization then proceeded as follows: in line with Twomey and Westermann (2017b), stimuli were presented in alternation for eight trials each. The familiarization phase therefore consisted of 16 trials in total. The initial stimulus was counterbalanced across simulations. In line with previous similar models, we used the network’s error on the output of the STM component as an index of infants’ looking times (Mareschal & French, 2000; Twomey & Westermann, 2017a; Westermann & Mareschal, 2012, 2014).

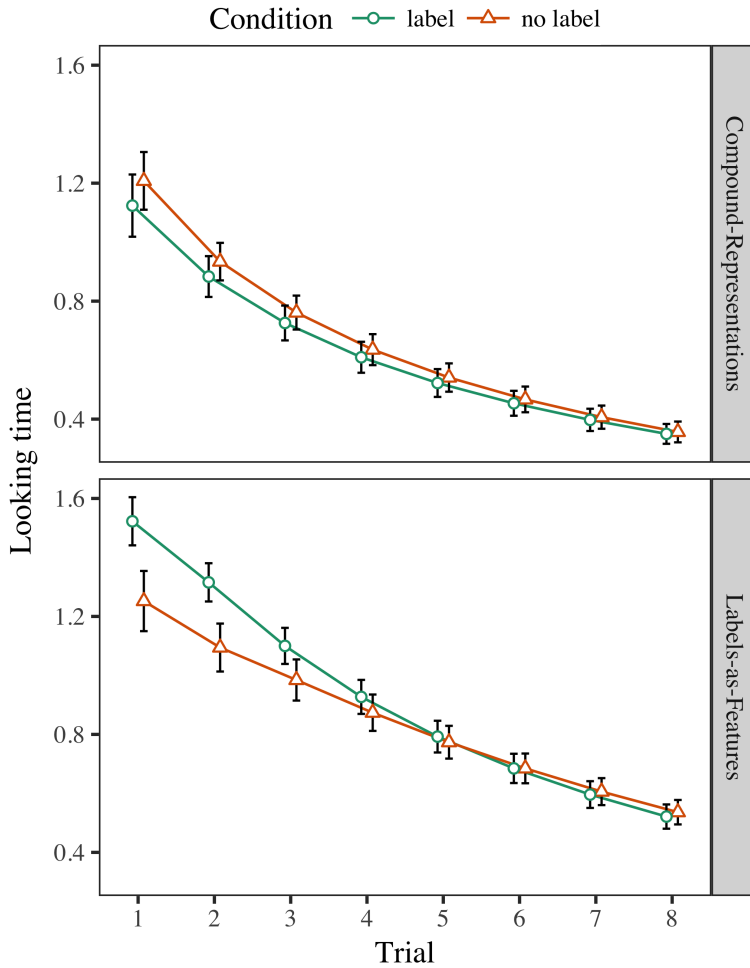


Figure 4: Looking time results for Experiment 1 simulations. Error bars represent 95% confidence intervals.

2.3 Results

Results from the familiarization phase for both simulations are depicted in Fig. 4. We submitted STM error (looking time) to an omnibus linear mixed-effects model using the R (3.4.4) package `lme4` (1.1-17) (Bates et al., 2015) (full code available on GitHub). The model with maximal random-effects structure that converged (Barr et al., 2013) included fixed effects for trial (1-8), theory (Compound-Representations, Labels-as-Features), condition (label, no label), and the trial-by-condition, theory-by-condition, trial-by-theory, and trial-by-theory-by-condition interactions; and by-subject random intercepts and slopes for trial and condition. The main effect of condition did not significantly improve model fit according to a likelihood ratio test; all other fixed effects analysis significantly improved model fit. Full details of the fitted fixed effect parameters and the likelihood ratio tests are provided in Table 1.

To understand the interactions, we submitted looking time for each model to separate mixed effects analyses, constructed in an identical fashion to the omnibus analysis. Full details of the theory-specific analyses' parameters are also given in Table 1. Overall, the CR model's looking time decreased rapidly across trials. There was a small but

Table 1: Estimated Parameters for Experiment 1 Looking Times: Fixed Effects for Global, CR, and LaF lmer Models

| Parameter | Global Model | | | |
|--|--------------|--------|---------|-------------|
| | Estimate | SE | X^2 | $Pr(> X^2)$ |
| Intercept | 0.998 | 0.0327 | | |
| Trial | -0.104 | 0.0038 | 270.735 | < .001 |
| Condition (no label) | 0.065 | 0.0230 | 3.185 | .074 |
| Trial \times Condition | -0.010 | 0.0040 | 30.712 | < .001 |
| Theory (LaF) | 0.434 | 0.0463 | 30.412 | < .001 |
| Theory \times Condition | -0.294 | 0.0325 | 16.000 | < .001 |
| Trial \times Theory | -0.039 | 0.0054 | 7.342 | .007 |
| Trial \times Theory \times Condition | 0.052 | 0.0056 | 82.828 | < .001 |
| Parameter | LaF Model | | | |
| | Estimate | SE | X^2 | $Pr(> X^2)$ |
| Intercept | 1.432 | 0.0312 | | |
| Trial | -0.143 | 0.0035 | 138.357 | < .001 |
| Condition (no label) | -0.229 | 0.0218 | 17.381 | < .001 |
| Trial \times Condition | 0.042 | 0.0036 | 119.769 | < .001 |
| Parameter | CR Model | | | |
| | Estimate | SE | X^2 | $Pr(> X^2)$ |
| Intercept | 0.998 | 0.0343 | | |
| Trial | -0.104 | 0.0041 | 128.776 | < .001 |
| Condition (no label) | 0.065 | 0.0242 | 2.549 | < .001 |
| Trial \times Condition | -0.010 | 0.0043 | 5.279 | .022 |

significant improvement in model fit; an interaction between trial and condition, with a slightly slower decrease in looking time in the label condition, but no main effect of condition. Thus, the CR model did not capture the pattern of results in the empirical study, in which infants looked longer at the previously labeled object. The LaF model's looking times also decreased across trials, and this model showed a strong effect of label, with longer looking times towards the previously labeled object. The trial-by-condition interaction also improved the model, with looking time towards the previously labeled object decreasing faster to fall to a comparable level to the looking time to the previously unlabeled stimulus. Although this interaction was not found in the empirical data analysis, it is not uncommon for models to deviate from the precise patterns of empirical data while capturing the overall pattern of interest. This is particularly the case with the additional noisiness found in infant data; the empirical data analysis might have failed to detect this interaction effect between trial and condition, due to the noisiness and smaller sample size of infant studies naturally decreasing statistical power. In the end, the LaF model captures Twomey and Westermann's (2017b) main empirical results of interest: when all else is held equal, teaching the LaF model a label for one object but not another leads to longer looking times towards the previously

labeled object in a subsequent, silent familiarization phase.

2.4 Discussion

In Experiment 1, we tested two possibilities for the relationship between labels and object representations using a neurocomputational model to capture recent empirical data (Twomey & Westermann, 2017b). The target data showed that previously learned labels affect 10-month-old infants' looking times in a silent familiarization phase, suggesting that knowing a label for an object directly affects its representation, even when that object is presented in silence. As noted by Twomey and Westermann (2017b), both the compound-representations (CR) and labels-as-features (LaF) accounts predict some effect of labels on object representations, and both theories could explain their empirical data. To disentangle these two accounts, we implemented both theories in simple dual-memory auto-encoder models inspired by Westermann and Mareschal (2014). In our CR model, we instantiated labels on the output layer only. This model learned to associate labels with inputs over time such that the presence of visual/haptic input for an object would consistently activate the label, but nonetheless, label information was separate from visual and haptic object information (Westermann & Mareschal, 2014). In our LaF model, labels were represented on the input as well as on the output layers in exactly the same way as the visual and haptic components of object representations (Gliozzi et al., 2009; Sloutsky & Fisher, 2004). Only the LaF model captured the longer looking to the previously labeled stimulus exhibited by the infants in Twomey and Westermann's (2017b) empirical study.

These results offer converging evidence that labels may have a low-level, featural status in infants' early representations. In line with recent computational work (Gliozzi et al., 2009; Westermann & Mareschal, 2014) we chose to explore such low-level accounts using a simple associative model that could account for the nuances of recent empirical data (Twomey & Westermann, 2017b). Our LaF model offers a parsimonious account of Twomey and Westermann's (2017b) results, in which looking time differences emerge from a low-level novelty effect (Sloutsky, 2003; Sloutsky & Fisher, 2004; Sloutsky et al., 2001), without the need to specify qualitatively different, top-down representations (Fulkerson & Waxman, 2007; Waxman & Booth, 2003; Waxman & Gelman, 2009). Specifically, as argued in Twomey and Westermann (2017b), and as implemented in the LaF model, over background training the label is learned as part of the object representation. Thus, when the object appears without the label there is a mismatch between representation and reality. This mismatch leads to an increase in network error for the previously labeled stimulus only, which has been interpreted in the literature as a model of longer looking times (Mareschal & French, 2000; Twomey & Westermann, 2017a; Westermann & Mareschal, 2012, 2014). Further, these results delineate between the two possible explanations for infants' behavior in the empirical task; specifically, our results support accounts of early word learning in which labels are initially encoded as low-level, perceptual features and integrated into object representations.

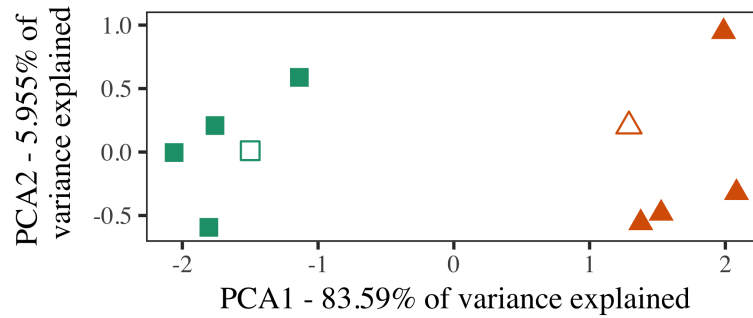


Figure 5: Example of two categories generated for Experiment 2 (first two dimensions of a PCA). Hollow shapes represent the prototypes, used during the familiarization (lab) phase, around which categories were constructed, and filled shapes represent exemplars used during background training. We used Principal Component Analysis (PCA) to reduce the dimensionality of the representational space in order to plot the 10-dimensional exemplars in a 2-dimensional space. The proportion of variance in the original representation explained by each of the plotted dimensions is specified on the axis labels.

3 Experiment 2

Overall, then, our LaF model offers a mechanism by which labels affect infants' representations of single objects. However, rather than one-to-one label-object mappings, infants typically learn labels for categories of objects; for example, a child might learn that their brown furry cuddly toy, the spotted animal in their picture book, and the hairy, barking animal at Grandma's are all referred to by the label 'dog'. A question that Twomey and Westermann's (2017b) empirical study and the current computational replication leave open, then, is whether the effect seen here would persist when considering richer categories rather than single objects. Thus, in Experiment 2 we extended our LaF model to category learning to make testable predictions for future empirical work. To this end, we trained our model with two object categories, one labeled and one unlabeled, before testing the model on a new exemplar from each category in the same way as in Experiment 1.

As our implementation of the CR model did not replicate the empirical results in Experiment 1, we did not use it in Experiment 2 and instead focused on the LaF model.

3.1 Stimuli

In these simulations, stimuli consisted of two distinct categories with five exemplars each. Four of the five exemplars for each category were used for background training, keeping the remaining one as a novel within-category item for the simulated looking time phase.

To allow for convenient future empirical testing of our predictions (e.g. using pictures in a storybook read at home as in Bornstein & Mash, 2010; Horst et al., 2011), we removed the haptic units from the model. We constructed our categories around two exemplars with one overlapping unit (out of the 10 visual units), and then randomly adding noise to this exemplar, adding to the prototype values taken from a uniform

Table 2: Estimated Parameters for Experiment 2 Looking Times: Fixed Effects for LaF lmer Model

| Parameter | Estimate | SE | X^2 | $Pr(> X^2)$ |
|--------------------------|----------|----------|---------|-------------|
| Intercept | 1.348 | 0.029841 | | |
| Trial | -0.153 | 0.0045 | 113.490 | < .001 |
| Condition (no label) | -0.350 | 0.0292 | 21.434 | < .001 |
| Trial \times Condition | 0.066 | 0.0052 | 138.707 | < .001 |

distribution between -0.5 and 0.5. Thus, we ensured that both categories formed distinct clusters in representational space, while making all exemplars within a category distinct from each other (Fig. 5).

3.2 Procedure

Similar to Experiment 1 we first trained the model with exemplars of each category, presented individually in alternating fashion, with timings drawn from a normal distribution of mean 2000 and standard deviation 200. Which category was labeled and which was unlabeled was counterbalanced across simulations.

We then presented the models with a familiarization phase in line with Experiment 1, in which the remaining exemplar for each category was presented without a label. As in Experiment 1, this phase consisted of 16 interleaved trials of up to 40 iterations (eight trials per category).

Again, to collect an amount of data consistent with infant studies, we ran a total of 40 model subjects.

3.3 Results

3.3.1 Looking Times

Using the same procedure as in Experiment 1, we fitted an omnibus linear mixed-effects model to the STM network error (looking time) during familiarization. Results are shown in Fig. 6. The final model included main effects of trial (1-8), condition (label, no label), and a trial-by-condition interaction; the model also included by-subject random intercepts, and random slopes for trial and condition. All fixed effects in this final analysis significantly improved model fit according to a likelihood ratio test. Full detail of the fitted fixed effect parameters are given in Table 2.

The model’s looking time decreased across trials (main effect of trial), and, as in Experiment 1, the model showed longer looking times towards the previously labeled category (main effect of condition), and a faster decrease in looking time towards this category (trial-by-condition interaction). Thus, the LaF model predicted that when trained with labeled and unlabeled categories rather than individual objects, infants should again show a novelty response when viewing silently-presented exemplars of the previously labeled category.

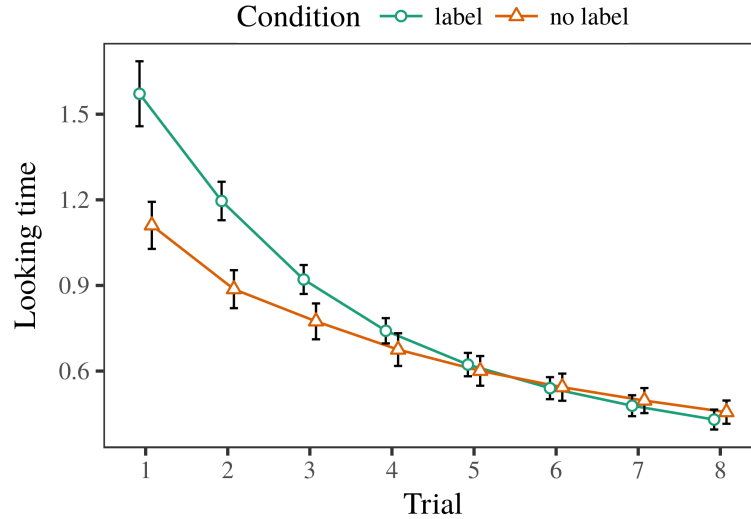


Figure 6: Looking time results for the Experiment 2 simulations. Error bars represent 95% confidence intervals.

Table 3: Parameters for Experiment 2 Internal Representations: Fixed Effects for LaF lmer Model

| Parameter | Estimate | SE | X^2 | $Pr(> X^2)$ |
|------------------------------------|------------|-----------|--------|-------------|
| Intercept | 1.635e-01 | 4.467e-03 | | |
| Step | 2.054e-03 | 1.321e-04 | 73.739 | < .001 |
| Condition (no label) | 1.815e-02 | 6.837e-03 | 4.891 | .027 |
| Step \times Condition (no label) | -2.752e-04 | 8.009e-05 | 11.774 | < .001 |

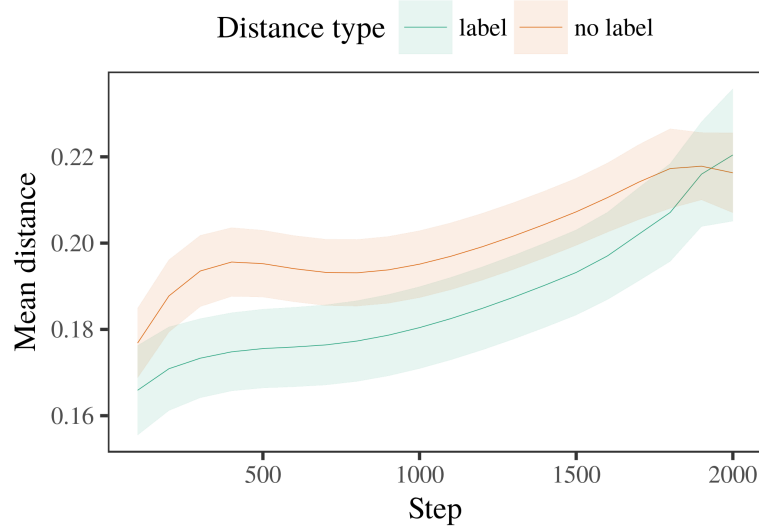


Figure 7: Evolution of mean distance in internal representations of the LTM during background training for Experiment 2 simulations. Shaded areas represent 95% confidence intervals.

3.3.2 Internal Representations in the Model

A common way to look at a neural network’s “understanding” of the inputs it has received is to examine the activation patterns in the hidden layer following encoding

(Mareschal & French, 2000; Rogers & McClelland, 2004; Westermann & Mareschal, 2012, 2014). We recorded these hidden representations for the training stimuli during background training every 100 iterations to investigate the development of memory representations. In our model, the LTM corresponds to representations in memory, whilst the STM corresponds to in-the-moment behaviors and perception; hence, we here examined the hidden units of the LTM network only. The mean within-category distances are displayed in Fig. 7.

We then submitted the mean distance between exemplars of each category to a mixed-effects model. We used the same model building principle as for the looking time results previously discussed.

The final model included main effects of step (iteration number when recording, divided by the recording interval of 100), a condition (label, no label), and a step-by-condition interaction; the model also included by-subject random intercepts and slopes for step and condition. All fixed effects in this final model significantly improved model fit according to a likelihood ratio test. The estimates for the fitted parameters of the fixed effects for this model are displayed in Table 3.

The mixed-effects model indicated that the within-category distance increased slowly over time (main effect of step), with the distances between exemplars of the unlabeled category being larger than the distances between exemplars of the labeled category (main effect of condition), and with distances in the unlabeled category growing more slowly than in the labeled category, after a quicker start (step-by-condition interaction). Thus, the presence of a label associated with a category in our LaF model caused exemplars of this category to be represented more closely together, and to be differentiated more slowly than in the unlabeled category.

3.4 Discussion

In Experiment 2 we extended our LaF model, which captured the empirical data from Twomey and Westermann (2017b) in Experiment 1, to a situation simulating infants' learning about object categories. The model predicted similar looking time patterns compared to those observed with single objects; that is, that infants should look longer, in silence, at exemplars that belong to a category for which they know a label.

Examination of the LaF network's hidden representations revealed that the labeled category was more compact than the unlabeled category, making labeled exemplars appear more similar to each other than unlabeled exemplars. The model nonetheless learned to discriminate different exemplars of a same category, making the distance between exemplars increase over time. The prediction that increased similarity between exemplars of a category may be seen together with longer looking times is intriguing. The reduced distances between exemplars of the labeled category in the model suggest that exemplars should be perceived as more similar to each other than those of the unlabeled category. If so, a new exemplar of this labeled category may be perceived as less novel than a new exemplar of the unlabeled category, leading to longer looking times to the latter. In contrast, however, the model predicts longer looking towards the previously

labeled category exemplar, despite the reduced distance in internal representations. Our interpretation of this counter-intuitive result is that, despite the labeled category being more compact, the surprise effect of seeing an exemplar of this category without a label is still stronger than the facilitatory effect of a reduced distance in representational space.

Notably, Westermann and Mareschal (W&M Westermann & Mareschal, 2014) used a CR model to address a related issue, specifically the effect of labeling on children’s longer-term category learning. In their model they found reduced looking times to novel category exemplars for which a label was known compared to those with an unknown label. The predictions made by our LaF model in Experiment 2 therefore diverge from those of W&M: although the LaF model, like W&M, predicted that a category label reduces within-category distance in mental representations, it predicted higher instead of lower looking times for novel label-known category exemplars.

The reason for this difference likely relates to differences in stimuli and training between W&M’s model and the current simulations. Specifically, W&M aimed more broadly to model the transition from prelinguistic to language based processing in infant development. W&M provided their model with a relatively rich background knowledge of 208 exemplars drawn from 26 real-world basic level categories from four superordinate categories that were encoded through 18 meaningful features (geometry, object characteristics). In their simulation of label effects on object familiarization, the model first received background training on 202 objects from all 26 categories, including two rabbits. In the no-label condition no objects were labeled, and in the label condition encountered objects were labeled half the time (accounting for the fact that objects are not reliably labeled at every instance in which infants experience them). Then, the models were familiarized on 6 novel rabbits. Under these circumstances, W&M found that the label model familiarized faster to these stimuli than the no-label model.

In contrast, here we aimed to predict a controlled lab experiment, which involves less naturalistic situations and stimuli, with a single age group. Thus our current model learned only two categories and saw a single test stimulus for each. During background training, objects from one of the categories were always labeled and objects from the other category were never labeled. Conversely, W&M’s categories were perceptually very broad, and overlapped with other categories. The introduction of labels in this environment warped the representational space so that overlapping representations became separated in accordance with the labels. In the simulations reported here, however, the two categories were tight and non-overlapping, so that the effects of labels were far more subtle. It is possible that the categories considered here are not sufficiently rich and variable for the label to become detached from each object’s featural representation across learning. Indeed, our categories are made of a handful of exemplars each, with a limited number of features with low variability defining their belonging to a category, which contrasts with real-world categories defined by more, and more variable features.

Finally, it may be the case that the effect of the label on infants’ category representations varies with age, perhaps developing from a labels-as-features representation to a compound-representations mechanism over time (Sloutsky et al., 2001). From this

perspective, our model may simulate an earlier developmental stage (and mechanism), than W&M. It is indeed possible that infants first perceive labels as object features and form categories purely on a similarity basis, then slowly learn that labels are highly reliable predictors of category membership, even for less perceptually similar objects (e.g. “furniture”, “animals”, or “toys” Sloutsky et al., 2001; Westermann & Mareschal, 2014). Empirical studies with infants are currently underway to address this issue.

4 General Discussion

The current simulations demonstrate that a labels-as-features account can explain empirical looking time data from 10-month-old infants pre-trained with one labeled and one unlabeled 3D object. Further, the LaF model predicted that when trained with labeled and unlabeled simple categories of objects, infants would exhibit longer looking times to a novel exemplar of the previously labeled category presented in silence. Testing this prediction experimentally is crucial; if confirmed, it would shed new light on categorization studies in infants, stressing that the same mechanisms (here compacting the representation of a category) might lead to very different, or even opposite behavioral results depending on the nature and structure of stimuli used.

It is important to note that other computational work has explored the effect of labeling on object representations in infants. Gliozzi et al. (2009) used a self-organizing map (SOM Kohonen, 1990) architecture to capture empirical data from a categorization task with 10-month-old children. Given that labels are represented as units in SOMs in the same way as visual features, this model might capture Twomey and Westermann’s (2017b) results for similar reasons to the success of the LaF model. However, the two networks make very different assumptions about learning mechanisms, highlighting an important issue for both infancy research and computational work. Gliozzi et al.’s (2009) model learns in an unsupervised way, strengthening associations between units in its SOM using “fire together, wire together” Hebbian learning. In contrast, our model learns by comparing what it “sees” to what it “knows” and updating its representations in proportion to any discrepancy. Thus, the current results are compatible with an error-based learning account to development, in which infants learn by tracking mismatches between representation and environment (Heyes, 2017). Whether unsupervised learning, error based learning, or some combination of both drives early development is a profound theoretical issue outside the scope of the current paper; for now, we highlight the importance of bearing in mind the link between the technical assumptions of a computational model and the implications for (developmental) theory.

In an era of increasing enthusiasm for complex, deep neural networks capable of learning to represent and label images, play (video) games, and many other tasks, it is important to show that simplicity in modeling can be a distinct strength. In particular, the simplicity of the architectures presented here produces a more transparent and interpretable mechanism than a network with many hidden layers. There would, however, be an obvious interest in the future in scaling up our work to increasingly complex – and therefore realistic – learning environments, ultimately taking our model from the

“friendly nursery” of our controlled setup and inputs into the real world. One important question is, for example, if a labels-as-features network would naturally evolve to give less and less importance to the input labels, effectively becoming a compound-representations model on the basis of experience with the world. This would support the hypothesis that infants learn through experience that labels are features with a higher predictive value for categorization, and therefore stop experiencing them as input features of object but learn to recall labels when presented with exemplar of known categories.

Finally, our simulations focused on two theories of the effect of labeling on category formation, but did not address the *labels-as-symbols* theory (Waxman & Markow, 1995). This theory assumes that labels are qualitatively different from other object features, and act in a symbolic way to directly shift the attentional focus towards diagnostic features that define a category. It is unclear how this theory could be implemented within the current framework, as our models do not have an explicit attentional component, and the very mechanism by which labels would highlight common features is not clearly defined in the theoretical account. Additional work is needed, on the one hand to define the precise mechanisms underlying this labels-as-symbols theory, and on the other hand to translate them into a computational model that can be tested and evaluated rigorously.

Taken together with Twomey and Westermann Twomey and Westermann (2017b), however, the current work demonstrates how language can shape object representation and in this way, explain empirical results in infancy research.

References

- Althaus, N., & Mareschal, D. (2014). Labels direct infants’ attention to commonalities during novel category learning. *PloS one*, *9*(7), e99670. <https://doi.org/10.1371/journal.pone.0099670>
- Althaus, N., & Plunkett, K. (2015). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, 1–11. <https://doi.org/10.1111/desc.12358>
- Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, *151*, 5–17. <https://doi.org/10.1016/j.jecp.2015.11.013>
- Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development*, *60*(2), 381. <https://doi.org/10.2307/1130984>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). <https://doi.org/10.18637/jss.v067.i01>
- Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, *81*(3), 884–897. <https://doi.org/10.1111/j.1467-8624.2010.01440.x>

- Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670. <https://doi.org/10.1126/science.146.3644.668>
- Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, *105*(1), 218–228. <https://doi.org/10.1016/j.cognition.2006.09.005>
- Gelman, S. A., & Coley, J. D. (1991). Language and categorization: The acquisition of natural kind terms. In *Perspectives on language and thought: Interrelations in development* (pp. 146–196). Cambridge, Cambridge University Press.
- Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, *22*(12), 2781–2789. <https://doi.org/10.1162/jocn.2010.21427>
- Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738. <https://doi.org/10.1111/j.1551-6709.2009.01026.x>
- Graham, S. A., & Poulin-Dubois, D. (1999). Infants' reliance on shape to generalize novel labels to animate and inanimate objects. *Journal of child language*, *26*(2), 295–320. <https://doi.org/10.1017/S0305000999003815>
- Heyes, C. (2017). When does social learning become cultural learning? *Developmental Science*, *20*(2), e12350. <https://doi.org/10.1111/desc.12350>
- Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*. <https://doi.org/10.3389/fpsyg.2011.00017>
- Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. <https://doi.org/10.1002/icd.364>
- Hurley, K. B., & Oakes, L. M. (2015). Experience and distribution of attention: Pet exposure and infants' scanning of animal images. *Journal of Cognition and Development*, *16*(1), 11–30. <https://doi.org/10.1080/15248372.2013.833922>
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. <https://doi.org/10.1109/5.58325>
- Mani, N., Durrant, S., & Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. *Journal of Memory and Language*, *66*(4), 612–622. <https://doi.org/10.1016/j.jml.2012.03.003>
- Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science*, *21*(7), 908–913. <https://doi.org/10.1177/0956797610373371>
- Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition*, *121*(2), 196–206. <https://doi.org/10.1016/j.cognition.2011.06.013>
- Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06

- McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1), 11–38. <https://doi.org/10.1111/j.1756-8765.2008.01003.x>
- Mirolli, M., & Parisi, D. (2005). Language as an aid to categorization: A neural network model of early language acquisition. In *Modeling language, cognition and action* (pp. 97–106).
- Morse, A. F., & Cangelosi, A. (2017). Why are there developmental stages in language learning? a developmental robotics model of language development. *Cognitive Science*, 41, 32–51. <https://doi.org/10.1111/cogs.12390>
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2), 665–681. <https://doi.org/10.1016/j.cognition.2007.04.003>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, 7(6), 246–251. [https://doi.org/10.1016/S1364-6613\(03\)00109-8](https://doi.org/10.1016/S1364-6613(03)00109-8)
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133(2), 166–188. <https://doi.org/10.1037/0096-3445.133.2.166>
- Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, 111(1), 65–86. <https://doi.org/10.1016/j.jecp.2011.07.007>
- Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child development*, 1695–1709. <https://doi.org/10.1111/1467-8624.00373>
- Twomey, K. E., & Westermann, G. (2017a). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, e12629. <https://doi.org/10.1111/desc.12629>
- Twomey, K. E., & Westermann, G. (2017b). Learned labels shape pre-speech infants’ object representations. *Infancy*, 23, 61–73. <https://doi.org/10.1111/inf.12201>
- Waxman, S. R., & Booth, A. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, 6(2), 128–135. <https://doi.org/10.1111/1467-7687.00262>
- Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, 13(6), 258–263. <https://doi.org/10.1016/j.tics.2009.03.006>

- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Westermann, G., & Mareschal, D. (2004). From parts to wholes: Mechanisms of development in infant visual object processing. *Infancy*, *5*(2), 131–151. https://doi.org/10.1207/s15327078in0502_2
- Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, *27*(4), 367–382. <https://doi.org/10.1016/j.cogdev.2012.08.004>
- Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391–20120391. <https://doi.org/10.1098/rstb.2012.0391>

Chapter 3

Unmatched Feature Saliency and Diagnosticity

In the previous chapter, we extended an existing neurocomputational model to account for the *labels-as-features* and *compound-representations* theories, and used it to test the predictions of these theories and their fit to existing empirical data from 10-month-old infants. Crucially, we provided evidence that the *labels-as-features* model was better able to explain the empirical results, thus suggesting that 10-month-old infants view labels in the same way as other features. We further used our model to predict results in an ongoing follow-up empirical study.

Now that we have proven the validity of our neurocomputational model, we set out to answer the main question we left open in our introductory chapter: how do visual feature saliency and auditory labels interact in shaping the way infants and adults explore new stimuli during a categorisation task? To answer this question, we plan to combine empirical and modelling results. We start in this chapter by presenting an empirical study on 15-month-old infants and adults.

Learning Categories with Conflicting Feature Salience and Diagnosticity

Arthur Capelier-Mourguy, Katherine E. Twomey, Gert Westermann
Department of Psychology, Lancaster University (UK)

Abstract

How do labels interact with objects in the process of category learning? This question has gathered a lot of interest, particularly in developmental psychology. Despite numerous studies and approaches, the role of labels in categorisation is still unclear; the *labels-as-features* theory argues that labels are first seen as one of the features of an object, whereas the *labels-as-symbols* theory considers labels to be referential for categories from an early developmental stage. Here we directly test the prediction of the latter theory that labels can drive attention to diagnostic features of objects. We do so by presenting 15-month-old infants and a control adult group, with categories where the salient feature (head) is non diagnostic, but a non-salient feature (tail) is diagnostic of category membership. According to the *labels-as-symbols* theory, we expected participants to shift their attention away from the salient head and towards the diagnostic tail when hearing category labels, compared to participants in a control condition. However we found that infants who heard a label did not significantly differ from infants in a control group during training, and still learned the label-category pairs as evidenced at test. These results provide indirect evidence against the *labels-as-symbols* theory, and most importantly, add to the converging evidence that looking behaviours are not a direct proxy for learning.

Keywords: development, labelling, categorisation, salience, diagnosticity

1 Introduction

Facing the complex world, infants have to bring the objects they encounter together into categories to make the world simpler and reduce the cognitive charge it requires to live in it. Infants automatically group together items that are similar, and separate items that are dissimilar, slowly building up categories based on what they see (Mareschal & French, 2000; Mareschal et al., 2000). Category exemplars are often encountered together with the name of the category spoken by a caregiver, and such naming events also have been argued to improve categorisation (e.g. Althaus & Westermann, 2016; Gelman & Coley, 1991; Gliga et al., 2010; Graham & Poulin-Dubois, 1999; Plunkett et al., 2008). However, the mechanism by which adding a spoken label improves categorisation processes in pre-linguistic children remains unclear.

A first theory suggests that labels are separate from object representations and act as referential pointers in a top-down way, inviting the listener to form categories (Waxman

& Markow, 1995). A possible mechanism for this theory is that labels drive attention towards diagnostic features, that is, features shared by all exemplars of the category but not by out-of-category items. For example, knowing that both giraffes and zebras have four legs and a long head is not helpful to discriminate them into two categories, while the neck length of giraffes and the stripes of zebras are both diagnostic features for their respective categories. Supporting this theory, studies have shown that adding a label specifically allows infants to form categories that they would not otherwise form, be it grouping two different sets of items into one category (Plunkett et al., 2008), or separating one category into two when exemplars have different labels (Althaus & Westermann, 2016). More recently, two eye-tracking studies showed that adding a label to a set of objects directed infants' attention towards features of low variability (Althaus & Mareschal, 2014), and increased their importance in the representation of the objects as shown by longer looking times towards those features of low variability in a subsequent test phase (Althaus & Plunkett, 2015a).

A second theory argues that labels are features, part of the object representation at the same level as other physical or auditory features: a dog is an animal with four legs, fur, a tail, a dog face, and is called “dog” (Sloutsky & Fisher, 2004). In this theory, labels simply facilitate categorisation by adding to the overall similarity of all exemplars within a category, since they all share the same name in addition to other features. In support of this theory is a study that contrasted categorisation and inference tasks (Deng & Sloutsky, 2012). In both tasks, participants were familiarised with a series of objects divided into labelled categories. A categorisation task then required participants, when presented with a single object, to deduce its category (i.e. the label that corresponded to the category), while an inference task required participants, when being given an incomplete object and a label, to infer the missing feature. If the labels act as a category markers, the expected behaviour in an inference task in which there is a mismatch between the label and other features is that the inferred feature would be of the category denoted by the label, and mismatching the other features—a result that was not observed. Instead, this study showed that a mismatched label did not override other features when inferring a missing feature.

Rather than having to choose between the two theories, there might be a developmental change from the latter to the former (Sloutsky et al., 2001). Evidence suggests that infants treat labels as features, relying on them only if no other highly salient and diagnostic feature is available, while adults rely more if not solely on labels to form categories, suggesting that they perceive labels as category markers. Nonetheless, adults might still, in some conditions at least, rely on labels as features rather than markers (Deng & Sloutsky, 2012). A third view introduced recently can account for this evolution in time (Westermann & Mareschal, 2014). This *compound-representations* account assumes that labels are encoded in the same representational space as other features, but are not integrated to the object representations, only linked to them. In this way, labels will first drive categorisation by adding to the within-category similarity. With learning, over time, labels will become more closely associated to object representations,

and act more like markers for categories, reducing the distance in representational space between exemplars of the same category.

Despite the numerous studies conducted on this topic, few have focused specifically on the online process of categorisation and effects of labels on this process. Rather, studies have often focused on the behavioural results of categorisation tasks, for example, successful categorisation or inference of a missing feature. Recently, some research has focused on the online process, recording participants' looking times towards different object features during categorisation in silent or in labelling conditions. In two studies with 8- and 12-month-old infants, labels were shown to drive attention to less variable features of a set of objects, deemed "diagnostic" (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015a). However only one abstract category was presented, and the experiment did not require participants to discriminate objects into two different categories. Thus, the presence of an actual categorisation mechanism is not clear, and the results could be explained by habituation to the different stimulus features and a novelty effect due to the presentation of highly contrasting stimulus features, rather than a more global object categorisation process. In another study using eye-tracking (Deng & Sloutsky, 2015), 8- and 12-month-old infants were presented with two labelled categories. Infants showed no difference in looking towards any of the features due to the presence of a label; in particular, the label did not drive attention to diagnostic features as predicted by the *labels-as-symbols* theory. However, in this study all features were equally diagnostic and thus there was no particular feature that the label could have highlighted, according to the *labels-as-symbols* theory.

A related study with 8-year-old children and adults aimed to ascertain whether labels help identification of similarities versus differences between category exemplars (Barnhart et al., 2018). Multiple contrasting conditions were presented within subjects: one of the categories had no label associated with it (control condition), a second had the same label for all exemplars (commonality condition), and a third had one specific label for each exemplar (unicity condition). The authors argued that in the commonality condition, the label should highlight commonalities between exemplars, when in the unicity condition, it should highlight differences between exemplars, and the results confirmed these predictions. However, it is possible that these two outcomes actually arise from the same, single mechanism: the label should highlight diagnostic features for a category, which means features that are both shared by all members of a category, and unseen in out-of-category objects. In the above-described experiment, the results could come from the fact that the commonality category was indeed formed as one category and the label highlighted the common feature shared by all category exemplars, when items in the unicity category were treated as belonging to many different sub-categories and the label highlighted the diagnostic features that made them different to other exemplars with different names.

Together, the existing research does not present a clear picture of how category labels affect the attention to and processing of object features, and in particular, how labels and feature salience interact in this process. In previous work, salience was either not

accounted for, or was controlled by making all features equally salient. In the present study we address the question of how category labels and feature salience interact in guiding attention to object features and in category learning. To our knowledge, no study has investigated whether or not a label can actively guide categorisation in categories where low-salience, but not high salience, features are diagnostic. Testing infants and adults allows us to look at how labels impact categorisation both early in development and in expert speakers, to test the hypothesis that labels might first act as features and take on a more referential role through development. In the current study, we presented participants with a series of simple snake-like animals with two features: a head (high salience) and a tail (low salience). Importantly, the high-salience head did not indicate category membership but varied pseudorandomly during familiarisation. In contrast, the low-salience tail was diagnostic of category membership. If labels drive attention towards diagnostic features during familiarisation with category exemplars, we expected that (a) participants who heard a label would, during the familiarisation phase, look more and/or more quickly at the tail, and encode it more robustly, and (b) participants who heard a label would form stronger categories. Additionally, in a subsequent novelty preference test on infants contrasting familiarised features with new features, we expected infants to exhibit preferential looking towards the new features only if they encoded enough information about the old features during the familiarisation phase (Houston-Price & Nakai, 2004).

2 Experiment 1

2.1 Methods

All materials used for this experiment are available online for inspection and replication purposes¹, including raw stimuli, the experiment script in Tobii Studio (version 2), raw data, and analysis scripts in R.

2.1.1 Data Handling and Software Specifications

Data Handling A common measure in eye-tracking data analysis is the proportion of looking at an area of interest (AOI). To account for the boundedness of proportion values, we used the arcsine-root transformation of the proportion in our statistical models; for ease of language, we use the term “proportion” to talk about this measure. However, we plot raw proportion values only, for ease of visual interpretation.

Further, we discarded looks outside of our defined AOIs. This means that, for example, the proportion of looking at the tail during the familiarisation trials is defined as the time spent looking at the tail divided by the time spent looking at either the tail or head, but not the total time spent looking at the screen during a trial.

Software Specifications All statistical results were obtained using R (version 3.6.1; R Core Team, 2019). Analyses in this paper were conducted using (a) `lme4` (version 1.1-

¹<https://osf.io/5yh67/>

17; Bates et al., 2015) to run Sample Theory Based (STB) (generalised) linear mixed-effects models, `lmerTest` (version 3.0-1; Kuznetsova et al., 2017) to run ANOVA analyses on those mixed-effect models, and `emmeans` (version 1.3.5.1; Lenth, 2019) to compute estimated marginal means for those mixed-effects models, (b) `eyetrackingR` (version 0.1.8; Dink & Ferguson, 2018) to handle eye-tracking data and run bootstrapped cluster-based permutation analyses, and (c) `ggplot` (version 2.2.1; Wickham, 2016) to plot graphs from our data and `ggeffects` (version 2.4.1; Lüdtke, 2018) to compute and plot estimated marginal effects from our models.

2.1.2 Participants

A total of 48 15-month-old infants (25 girls, $M = 451.8$ days, range 430-469 days) provided data for the study. A further 17 participants were excluded for not meeting our inclusion criteria (minimum 50% of looking on 50% of the familiarisation trials, $n = 16$), or technical error ($n = 1$). Infants were randomly assigned to the label ($n = 24$) or the no-label ($n = 24$) condition. All participants were English-learning monolingual infants with no reported history of developmental delay.

2.1.3 Materials

Visual Stimuli Based on previous studies which demonstrate that infants have a strong bias for looking at heads (Quinn et al., 2009), our stimuli consisted of simple snake-like animals with two features only (a head and a tail) to ensure stimuli afforded a “natural” non-uniform salience that all participants would share. Each feature varied around two prototypes, and stimuli were created so that the head and tail dimensions would be independent.

In the label condition, we defined two categories, accounted for by the low-salience tail. Therefore, the tail varied consistently with the label, being fully diagnostic of the category, while the head varied pseudorandomly with the label (i.e. there were multiple heads of each type associated with each label).

Fig. 1 shows an example of a familiarisation stimulus. Fig. 2 displays all stimuli used during familiarisation as pairs of features, with the horizontal line dividing the stimuli into the labelled categories for this study, and the vertical dashed line dividing the two categories depending on the two kinds of head. Fig. 3 displays test stimuli as pairs of features.

Auditory Stimuli During the first half of the familiarisation phase, the carrier phrase for the label/pronoun was “Look at [this]!” (no-label condition) or “Look at [the Saldie/the Gato]!” (label condition), then “Can you see [this]?” (no-label condition) or “Can you see [the Saldie/the Gato]?” (label condition) in the second half of the trials. All phrases were recorded in infant-directed speech by a female native English speaker.

The labels “Gato” and “Saldie” were chosen to be phonetically plausible English words that were not used in any previous studies. Those two words are actual words in

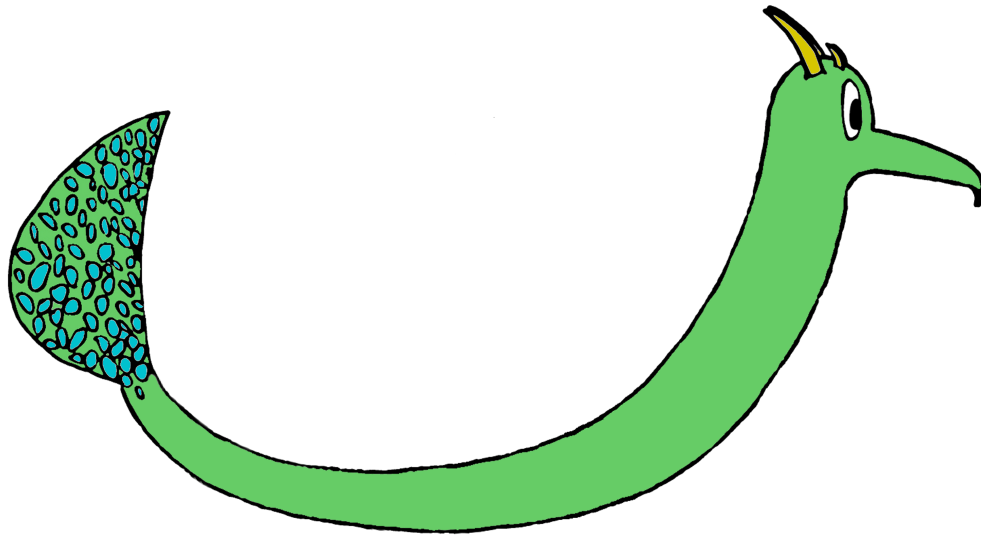


Figure 1: Example of a stimulus used for categorisation.

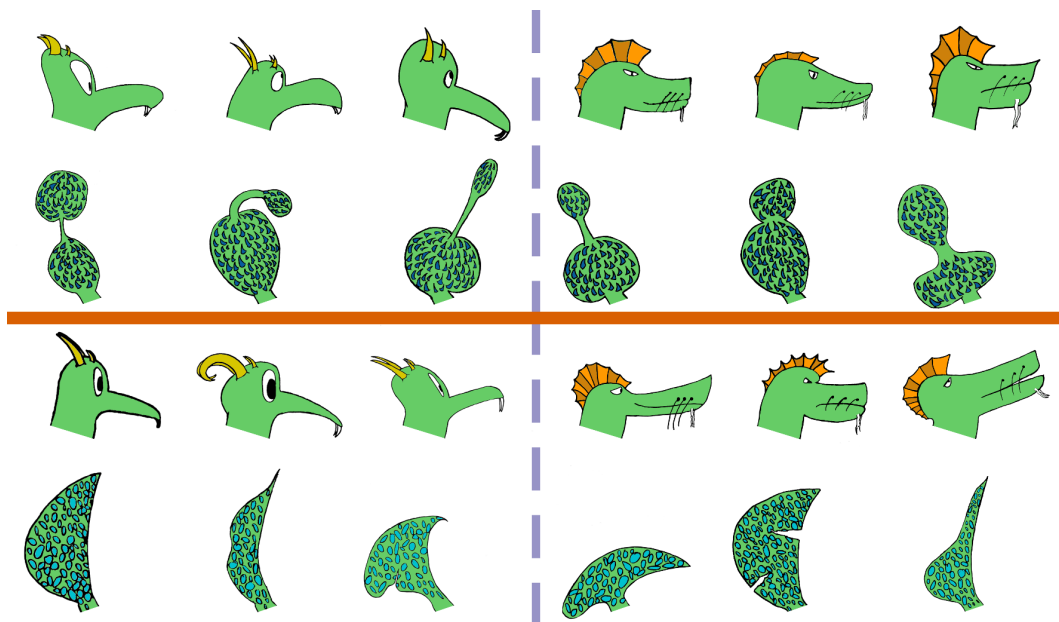


Figure 2: Pairs of visual stimuli used for familiarisation, grouped by tail type (horizontal line) and head type (vertical dashed line).

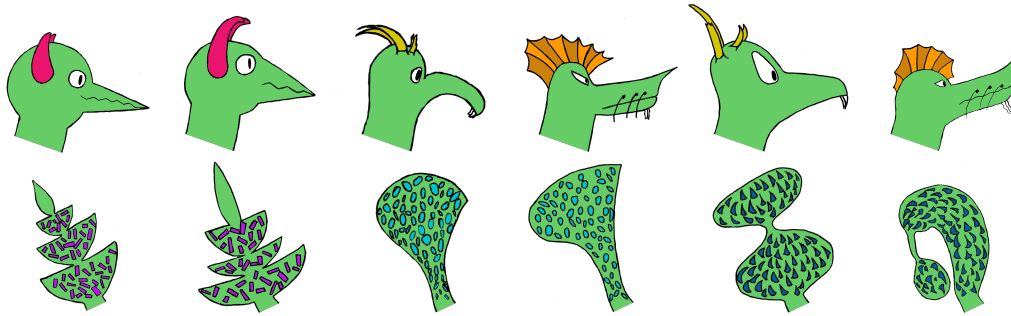


Figure 3: Pairs of visual stimuli used for contrast (and word-learning) tests. New features on the left, old features on the right.

Basque², but this was not an issue since our sample consisted of monolingual English infants only.

The volume for all recordings was normalised and further manually adjusted to obtain equivalent hearing levels.

2.1.4 Procedure and Design

During the experiment, infants sat on their caregiver’s lap approximately 60cm from a 23 inch, 1920x1080 pixels presentation screen. Eye-tracking data were collected using a Tobii X120 eye-tracker, calibrated using a child-friendly 9-point routine. No infant had to go through the calibration phase more than once.

After calibration, the experiment began with a familiarisation phase consisting of 24 trials. Each trial started with an animal spiralling towards the centre of the screen in silence from a top corner for 1500 ms and jiggling in the centre of the screen for a further 1000 ms to capture infants’ attention. Next, the animal stopped moving and a carrier phrase started for 1500 ms. The asynchronous presentation of the visual and auditory stimuli was important as synchronous presentation can lead to an auditory overshadowing effect: a preferential processing of the auditory signal over the visual signal, rather than a processing of both auditory and visual information and integration of those modalities together (Althaus & Plunkett, 2015b). The animal then remained still for another 3000 ms, until it slid away to a bottom corner on the last 500 ms of the trial. A full trial thus lasted for 7500 ms.

Successive trials presented animals belonging to alternating categories, and the first category presented was counterbalanced between infants. The matching of a label to a specific category was also counterbalanced between infants in the label condition. All six animals from each category were presented in a random order, and this presentation of 12 animals was repeated twice, leading to the full 24-trials familiarisation phase. The direction that each animal was facing was randomised, and all the movements were made so that the animal would be moving forward (e.g. an animal facing right would spiral in from the top-left corner and slide out to the bottom-right corner). An attention getter

²for our keenest readers, the original words were *gatu* meaning “cat”, and *zaldi* meaning “horse”

was presented after 8 and 16 trials.

Following familiarisation, the extent to which infants had encoded both features was then measured using a preferential-looking paradigm. Infants saw “contrast” test trials in which two animals were presented side by side, one with new exemplars of the features presented during the familiarisation phase (hereafter “old features”), and one with an old feature and an entirely new feature.

In total, there were three such contrast test trials: (a) a head contrast, presenting an animal with an old tail and head next to an animal with an old tail and a new head, (b) a tail contrast, presenting an animal with an old tail and head next to an animal with a new tail and an old head, and (c) a relative contrast, presenting an animal with an old tail and a new head next to an animal with a new tail and an old head. The order in which the head and tail contrast trials were presented was counterbalanced between infants, but the relative contrast was always shown last. An attention getter was presented before each test trial to ensure infants were fixating centrally before the onset of the trial, and the animals were always arranged so that the new feature would be at the side of the display, not in the centre.

Finally, infants in the label condition were presented with four word-learning test trials, in which they saw two animals side by side and heard “Look at the [Saldie/Gatoo]!”, with label alternating between trials. Both types of old heads and old tails were presented on the screen for each trial, and those were the same as the old features that were presented during contrast tests (i.e. old features that were not presented during familiarisation). The naming order, horizontal arrangement, and facing of animals were counterbalanced between infants. An attention getter was presented before each trial.

All test trials lasted for ten seconds or until infants looked away for more than two seconds as judged by the experimenter.

2.2 Results

Analysis Structure We conducted two types of analysis in this report: testing average proportion looking during one or several time windows of a trial, and time-course analysis. We also tested for other unique-per-trial values, however these tests followed the same structure as tests on proportion looking.

For the tests of proportion looking, we used (generalised) linear mixed-effects regression models fitted with maximal converging random-effects structure to estimate parameters (Barr et al., 2013). For significance testing of those parameters, we used type I ANOVA analyses with Satterthwaite’s method as implemented in `lmerTest` for linear models, and commonly-used asymptotic Wald tests for generalised linear models.

For the time-course analyses, we used bootstrapped cluster-based permutation analysis as implemented in `eyetrackingR` with 100 ms time bins and *t*-test comparisons between the two conditions (no-label, label); the choice of a *t*-test rather than a mixed-effects model was due to the current implementation in `eyetrackingR` that did not allow for the use of mixed-effects models when testing a between-subject factor as in our case. To test different levels of other factors (e.g. first three trials against last three trials),

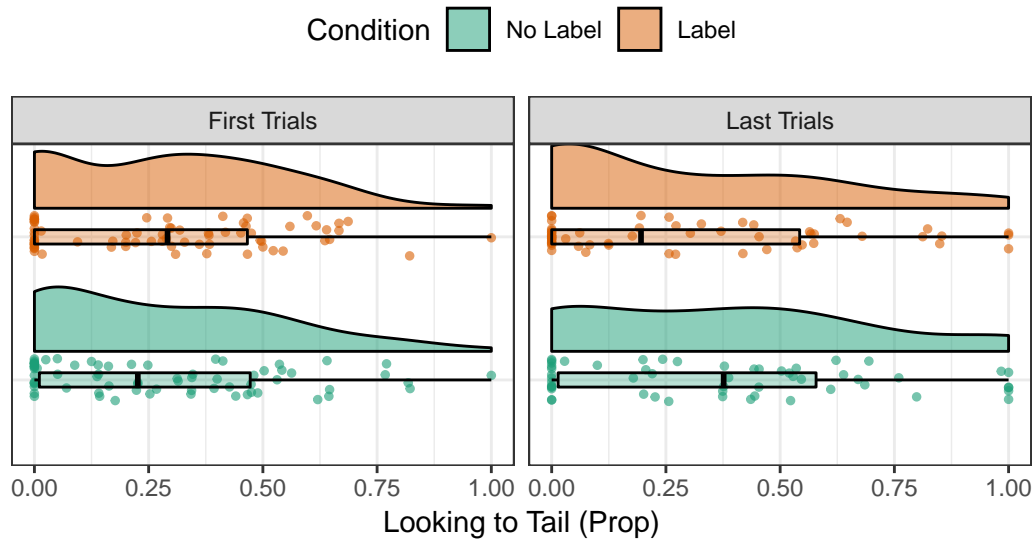


Figure 4: Raincloud plot from the data of the proportion of looking at the tail after label onset.

| Parameter | Model Output | | ANOVA Output | |
|-------------------|--------------|------------|--------------|-----------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 0.48 | 0.06 | | |
| FstLstLast Trials | 0.13 | 0.09 | 1.30 | 0.26 |
| ConditionLabel | 0.01 | 0.08 | 0.58 | 0.45 |
| FstLst:Condition | -0.12 | 0.12 | 0.96 | 0.33 |

Table 1: Parameter estimates and ANOVA results for the STB model on proportion looking at the tail after label onset.

we simply ran an independent analysis on each level of this factor (or levels of their interaction when using multiple factors); although this approach involved multiple comparisons, there was to our knowledge no straightforward way to test for multiple factors directly.

2.2.1 Familiarisation

Proportion of Tail Looks We submitted proportion looking at the tail during the 3000 ms following label onset to a linear mixed-effects regression model. The model included main effects of and interaction between the first/last three trials of the experiment per participant (FstLst), and Condition (no-label, label). The model also included random intercepts and slopes for FstLst by participant, and random intercepts by visual stimulus. A summary of the model’s parameter estimates and ANOVA results for those parameters are given in Table 1. A “raincloud” plot (Allen et al., 2019) of the data is shown in Fig. 4. These plots include half a violin plot to understand the shape of the data, individual data points to better understand the structure of the data, and a boxplot to give some descriptive statistics at a glance.

Notably, none of the parameters reached significance: there was no evidence for a difference in looking at the tail between the first few and last few trials in the no-label

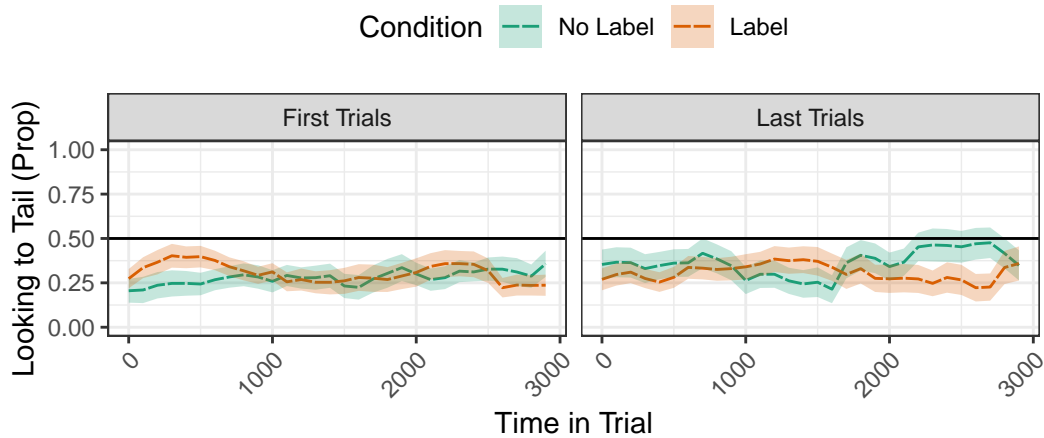


Figure 5: Time-course plot of the mean and SE of proportion looking at the tail.

| Parameter | Model Output | | ANOVA Output | |
|-------------------|--------------|------------|--------------|-----------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 6.56 | 0.21 | | |
| FstLstLast Trials | 0.30 | 0.31 | 2.28 | 0.13 |
| ConditionLabel | -0.34 | 0.29 | 1.80 | 0.19 |
| FstLst:Condition | 0.05 | 0.44 | 0.02 | 0.90 |

Table 2: Parameter estimates and ANOVA results for the STB model on first tail (AOI) hit during familiarisation.

condition, and no evidence for a difference between infants in the no-label and label condition at any point. In short, we cannot draw any strong conclusions from these results.

Time-Course Analysis Testing more finely for differences in proportion looking at the tail between conditions during the course of trials, we ran one bootstrapped cluster-based permutation analysis each for the first three and last three trials. The data are displayed in Fig. 5. No clusters were found to differ significantly from the null hypothesis: at no point within the first three or last three trials did infants in the label condition look significantly differently to the tail from infants in the no-label condition.

First Tail Look Another hypothesis we formulated was that infants would look more quickly at the diagnostic feature when hearing a label compared to infants who did not hear a label. To test this, we submitted the log-transformed time to first look at the tail after the animal had stopped moving in to a linear mixed-effects regression model. The model included main effects of and interactions between FstLst (first trials, last trials) and Condition (no-label, label), as well as random intercepts and FstLst slopes by participant. The model’s parameter estimates and ANOVA results for those parameters are given in Table 2. A raincloud plot of the data is shown in Fig. 6.

Here again, no effect was found to be significant: there was no evidence for a difference in time to first look at the tail between the first few and last few trials in the no-label

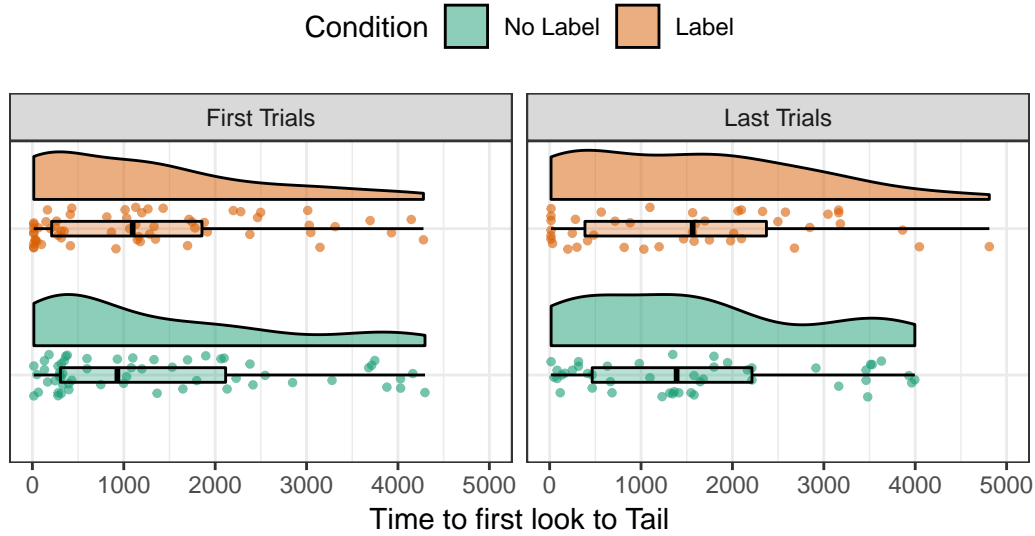


Figure 6: Raincloud plot from the data of the time before first look at the tail.

condition, and no evidence for a difference between infants in the no-label and label condition at any point.

2.2.2 Contrast Tests

Old-New Feature To test whether or not infants had encoded information from the tails and heads they had been familiarised with, we focused our analysis on the head-contrast and tail-contrast test trials. Out of the 48 participants who were included based on our criteria on the familiarisation trials, only 47 infants provided data for at least one trial, based on our per-trial inclusion criteria of looking at the screen 50% of the time: 23 in the no-label condition (of whom nine only contributed to the head contrast trial), and 24 in the label condition (of whom eight only contributed to the head contrast trial).

In this analysis, we considered two AOIs only, old and new feature, leaving out the centre of the screen which depicts two old features (either two heads in the tail-contrast, or two tails in the head-contrast). We then tested the proportion of looking at the new feature against chance, as a measure of novelty preference.

We submitted the chance-corrected proportion looking at the new feature to a linear mixed-effects regression model. The model included main effects of and interaction between ContrastType (head contrast, tail contrast) and Condition (no-label, label). The model further included random intercepts and slopes for ContrastType by participant. However, we were here interested in knowing whether or not the average looking time to the new feature in each group was significantly above chance. We thus report estimated marginal means (EM means) for the model and Bonferroni-corrected p -values in Table 3. A raincloud plot of the data can be seen in Fig. 7.

This post-hoc analysis suggests that only infants in the no-label condition had a strong preference for the new tail. No other results were significant: we have no evidence of a novelty preference for infants in the label condition, or for infants in the no-label

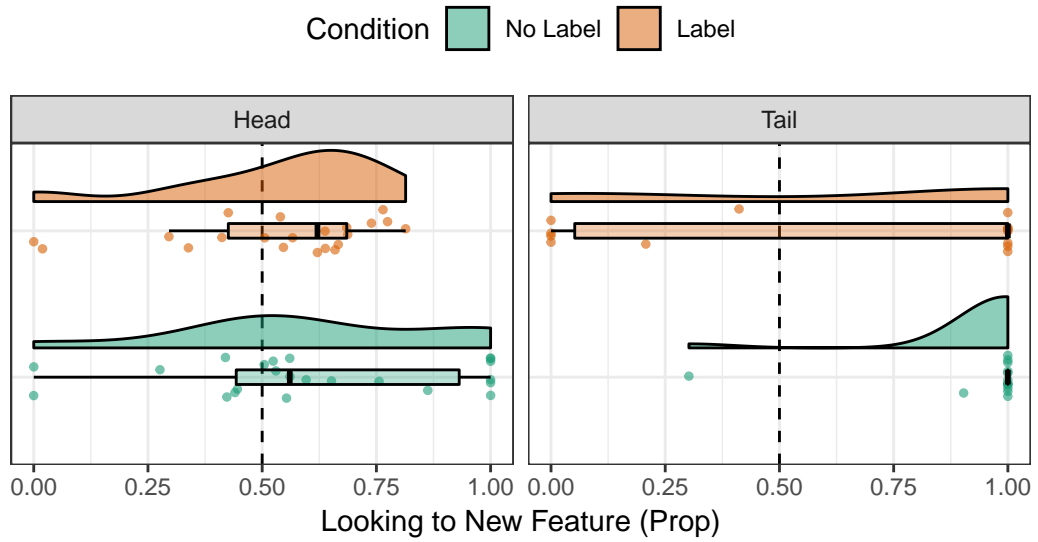


Figure 7: Raincloud plot from the data of the proportion of looking at the new feature.

| Group | EM Mean | 95% CI | t | $Pr(> t)$ |
|----------|---------|---------------|------|-------------|
| No Label | | | | |
| Head | 0.17 | [-0.05, 0.38] | 1.73 | 0.18 |
| Tail | 0.69 | [0.41, 0.97] | 5.66 | < .000 |
| Label | | | | |
| Head | 0.02 | [-0.21, 0.22] | 0.22 | 1.00 |
| Tail | 0.20 | [-0.08, 0.48] | 1.60 | 0.23 |

Table 3: Estimated Marginal (EM) means per group and Bonferroni-corrected p -values for the STB model.

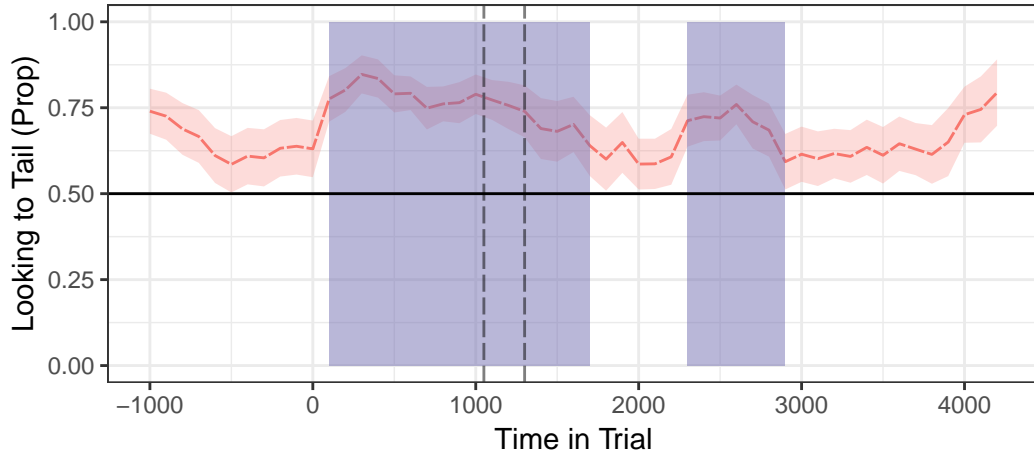


Figure 8: Time-course plot of the mean and SE of proportion looking at the target. Carrier phrase starts at $t = 0$, vertical dashed lines mark label onset for the two carrier phrases.

condition during head contrast trials.

However, we can see from the individual data points in the label-tail subplot that the null result there does not reflect our data: rather than having infants in the label condition all displaying equal looking to both the old and new tail, we clearly see a bimodal distribution with most infants having either a strong preference for the old or new tail. In other words, some infants exhibited a novelty preference ($n = 8$), while others exhibited a familiarity preference ($n = 4$), and only two infants had no strong preference.

2.2.3 Word Learning Tests

Next, we tested whether infants in the label condition had learned to match each label to its corresponding category. To do so, we considered two AOIs only: the target animal (with a tail matching the label), and the distractor animal (with the opposite set of features). We then conducted a bootstrapped cluster-based permutation analysis on the chance-corrected proportion of looking at the target over the course of the test trials. A time-course plot of the data and clusters of significant difference from chance can be seen in Fig. 8.

For this analysis, we chose t_0 as being the carrier phrase onset rather than the label onset. This choice was made post-hoc when noticing that infants started looking at the target before labelling but after the carrier phrase started, which can be explained by the fact that the carrier phrase for each label had slightly different phonetic properties. Thus, while we cannot be sure that infants learned the labels matching each category, we have evidence that infants matched the auditory cues in the carrier phrase to the corresponding categories.

Finding that infants succeeded in learning these categories when we did not find any evidence for differences in the familiarisation phase in terms of looking proportion and first look to the diagnostic tail, or in the contrast test trials in terms of looking

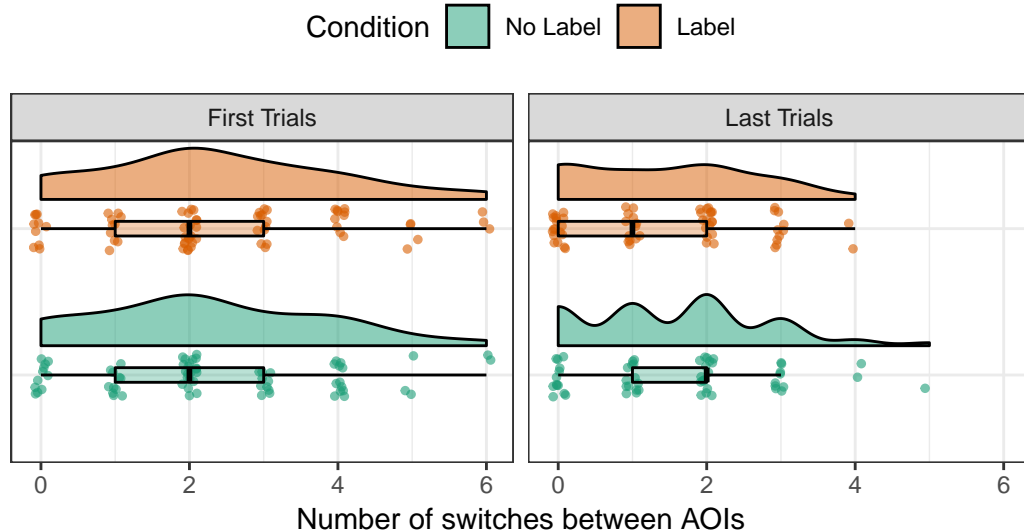


Figure 9: Raincloud plot from the data of the number of switches between AOIs.

more at a new tail, raised the question of how infants demonstrated learning on the word-learning test trials despite a lack of evidence for learning elsewhere. We therefore conducted further post hoc analyses on the familiarisation trials: number of switches between AOIs, and first AOI looked at.

2.2.4 Additional Analyses

Number of Switches The number of switches between different features of an object is commonly seen as a measure of distributed attention, which is believed to be linked to better encoding (e.g. Bronson, 1991; Colombo et al., 1991; Jankowski et al., 2001; Rose et al., 2003). Thus, a higher number of switches between the head and tail in the label condition could explain how these infants managed to learn the label-category match.

We submitted the number of switches between AOIs per trial to a Poisson linear mixed-effects regression model. The model included main effects of and interactions between FstLst (first trials, last trials) and Condition (no-label, label), as well as random intercepts and FstLst slopes by participant. The model's parameter estimates and p -values for those estimates can be found in Table 4. A raincloud plot of the data can be seen in Fig. 9.

These results suggest that infants in the no-label condition made significantly fewer switches on the last trials than on the first trials, and infants in the label condition did not differ significantly from infants in the no-label condition during the first or last trials.

First AOI Look We submitted the first AOI (head or tail) looked at after the animal had stopped moving in to a binomial linear mixed-effects regression model. The model included main effects of and interactions between FstLst (first trials, last trials) and Condition (no-label, label), as well as random intercepts and FstLst slopes by participant. The model's parameter estimates and p -values for those estimates can be found

| Parameter | Model Output | | ANOVA Output | |
|-------------------|--------------|------------|--------------|-------------|
| | Estimate | Std. Error | z value | $Pr(> z)$ |
| (Intercept) | 0.77 | 0.10 | 7.61 | 0.000 |
| FstLstLast Trials | -0.37 | 0.14 | -2.67 | 0.01 |
| ConditionLabel | 0.06 | 0.14 | 0.43 | 0.66 |
| FstLst:Condition | -0.16 | 0.19 | -0.84 | 0.40 |

Table 4: Parameter estimates and ANOVA results for the STB model on number of switches between AOIs during familiarisation.

| Parameter | Model Output | | ANOVA Output | |
|-------------------|--------------|------------|--------------|-------------|
| | Estimate | Std. Error | z value | $Pr(> z)$ |
| (Intercept) | -1.74 | 0.35 | -4.93 | 0.000 |
| FstLstLast Trials | 0.04 | 0.62 | 0.07 | 0.95 |
| ConditionLabel | 0.47 | 0.45 | 1.02 | 0.31 |
| FstLst:Condition | -0.09 | 0.73 | -0.13 | 0.90 |

Table 5: Parameter estimates and ANOVA results for the STB model on first AOI hit during familiarisation.

in Table 5. A histogram of the data can be seen in Fig. 10.

These results suggest that infants in the no-label condition looked significantly more often first at the head during the first trials than during the last trials, but no other difference was significant.

2.3 Discussion

In Experiment 1, we explored the effect of labelling on object perception and category encoding in 15-month-old infants. Specifically, we addressed the question of changes in attention distribution in the presence of a label when this label was presented with a diagnostic non-salient feature (tail) and a non-diagnostic salient feature (head). We found evidence that infants in the label condition learned to match categories to auditory stimuli, but no difference between conditions in terms of looking behaviour during familiarisation, and no clear evidence of familiarisation to either feature for all infants in the label condition as would be shown by a novelty preference. Critically, this result does not replicate results from the literature around the question of the effect of labelling on categorisation and perception in infants.

There could be two explanations for this lack of evidence: either adding a label truly does not impact looking behaviour in the setup we used, or there is a significant difference but we lacked statistical power to detect it. Although meta-analyses suggest that the latter is probably true (e.g. Lewis et al., 2016), it is interesting to consider the implications of the former.

A first possible explanation of the results we observe here is that of a ceiling effect, in which the head preference was so strong in both conditions that adding a label associated solely with the tail was not enough to direct infants' attention towards the diagnostic tail. To control for this possible effect, future work should seek to reduce the difference in salience between the different features of an object.

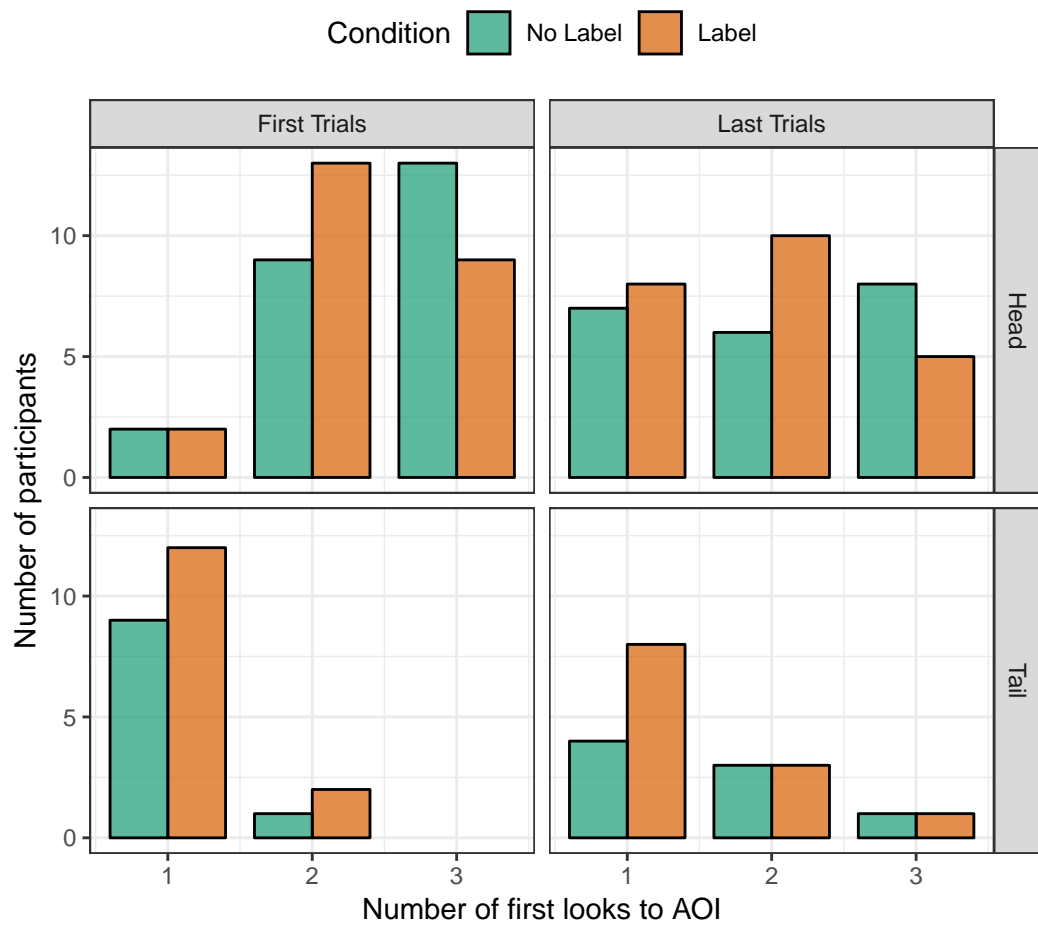


Figure 10: Histogram from the data of first AOI looked at.

Alternatively, the issue may be methodological. Historically in infancy research it has been assumed that eye-tracking measures are a proxy for information processing. However it arises from this study and previous work that information processing, as exhibited by evidence of learning at test, can happen without showing as systematic patterns of looking during training (e.g. Aslin, 2007; Hilton et al., 2019; Hilton & Westermann, 2017; Twomey et al., 2018). In particular, in our study, since infants in the label condition processed the information they needed to form the correct categories without looking longer or differently at the diagnostic tail, compared to infants in the no-label condition, the extent to which eye-tracking indexes information processing is unclear.

Nonetheless, infants in the label condition did show a preference in the tail contrast trials: some of them preferred the new tail as we expected, when others preferred the old tail, which is commonly seen as a need for further processing of the stimulus (Houston-Price & Nakai, 2004). We further looked into how these two groups of infants might vary in terms of word learning scores or looking behaviour during familiarisation. However the reduced sample size did not allow us to run reasonable statistical tests. Unreported diagnostic plots do not, however, suggest that infants who exhibited a familiarity preference looked less at the target during word learning trials, as could have been expected.

Overall then, it is unclear how labels affect information processing in this categorisation task in 15-month-old infants; however our data are more compatible with the *labels-as-features* theory, which predicts no attentional effects of labelling, than with the *labels-as-symbols* theory, which predicts that labels actively drive attention to diagnostic features. In Experiment 2, we asked whether in adults, labels would play a referential role in contrast to the featural role predicted in infants. To be able to link and compare results from both experiments, we presented adult participants with an explicit categorisation task, using the same material as in Experiment 1.

3 Experiment 2

3.1 Methods

All materials used for this experiment are available online for inspection and replication purposes alongside materials from Experiment 1, including raw stimuli, the experiment script in Eprime (version 2), raw data, and statistical scripts in R.

3.1.1 Data Handling and Software Specifications

The data handling and software specifications for this experiment are the same as for Experiment 1.

3.1.2 Participants

We recruited 60 participants from Lancaster University via an online pool of participants for psychology studies. Most participants were students, some of them studying

psychology. After exclusion of four outliers (more than two standard deviations away from the mean) in terms of learning speed, the final sample consisted of 56 participants (42 female, $M_{age} = 21.14$, range 18-39). All participants were fluent in English.

3.1.3 Materials

Visual Stimuli We used the exact same visual stimuli as in Experiment 1, since as for infants, the head is more salient for adults in animal-like stimuli (Kovic et al., 2009).

Auditory Stimuli After categorising each exemplar, participants in both conditions were given auditory feedback in the form of a shimmering sound for correct or a buzzer for incorrect categorisation. Then, participants in the label condition heard the phrase “It’s a [Saldie/Gatoo]”, pronounced by a female native British speaker in a neutral tone. The duration of both feedback sounds and both labelling phrases was the same.

3.1.4 Procedure and Design

Participants were tested in a quiet room, using a Tobii X120 eye-tracker calibrated using a 9-point routine to record eye-tracking data, and Eprime to run the experiment and collect behavioural data (categorisation responses, reaction time, number of training blocks, etc.).

The experiment consisted of a categorisation task: participants were presented with one exemplar at a time, and were asked to sort them into one of two categories by pressing the corresponding button on a keyboard. Participants were first presented with a training phase, during which they were provided with feedback after each categorisation decision. Participants in all groups heard non-linguistic feedback, followed by the label for the category for half of the participants. This training phase lasted for up to 21 blocks, or until successful categorisation (i.e. one full block without any mistakes). Each block consisted of the 12 exemplars shown in Fig. 2 presented in a random order. A fixation cross was presented in the middle of the screen for one second before each trial.

Participants were then presented with a test phase that consisted of the same categorisation task for one block without feedback. Two of the old exemplars for each category were replaced with new exemplars, to control for rote learning of category information for each exemplar separately, rather than formation of a feature-defined category.

3.2 Results

We used the same analysis structure for adults as we did for infants.

3.2.1 Behavioural Results

We hypothesised that participants’ categorisation abilities would benefit from hearing a label on top of the ‘correct/wrong’ auditory feedback. This outcome would be reflected

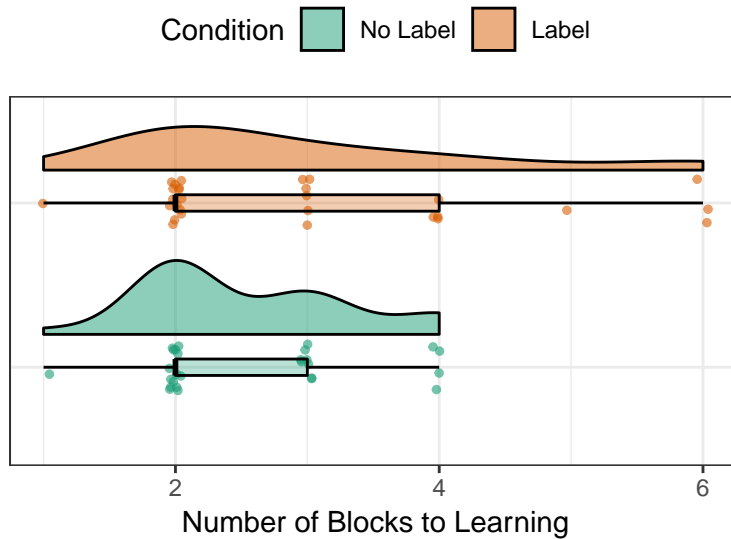


Figure 11: Raincloud plot from the data of the number of blocks to learning.

in the number of training blocks they needed to learn the categories (i.e. complete a full block without any categorisation mistakes), as well as in the overall response accuracy throughout training.

Number of Blocks to Learning Given that we did not have enough data points for a mixed-effects model and that this measure was not normally distributed (Anderson-Darling normality test: $A = 4.49$, $p < .001$), we conducted an independent 2-group Mann-Whitney U test to test the effect of labelling on the number of training blocks to learning. We found no significant difference between the label and no-label group ($W = 348.50$, $p = .45$). Notably, the median in both groups was of two blocks to learning, suggesting a ceiling effect (two blocks being the minimum number of blocks assuming that the participants infer the category structure after a few mistakes in the first block and then complete the second block with no mistakes).

Accuracy During Training We submitted accuracy during training to a binomial linear mixed-effects regression model. The model included main effects of and interaction between Condition (no-label, label), scaled log reaction time (zLogRT), and Block (starting at 0). The model also included random intercepts and slopes for zLogRT and Block by participant, and random intercepts by visual stimulus and by auditory stimulus. A summary of the parameter estimates and p -values for this model can be found in Table 6.

Participants in the no-label condition performed above chance in the first block (significant Intercept), their performance increased throughout training (significant effect of Block), and finally, participants in the no-label condition with longer reaction times later in training were also less accurate (significant effect of zLogRT:Block). No other effects were significant; particularly, participants in the label condition did not differ significantly from participants in the no-label condition at any point or in terms of

| Parameter | Estimate | Std. Error | z value | $Pr(> z)$ |
|----------------------------|----------|------------|-----------|-------------|
| (Intercept) | 0.77 | 0.25 | 3.11 | 0.002 |
| ConditionLabel | 0.13 | 0.32 | 0.41 | 0.68 |
| z LogRT | -0.12 | 0.16 | -0.72 | 0.47 |
| Block | 1.96 | 0.31 | 6.40 | < .001 |
| Condition: z LogRT | -0.25 | 0.22 | -1.11 | 0.27 |
| Condition:Block | -0.45 | 0.28 | -1.59 | 0.11 |
| z LogRT:Block | -0.61 | 0.19 | -3.27 | 0.001 |
| Condition:Block: z LogRT | 0.27 | 0.20 | 1.37 | 0.17 |

Table 6: Parameter estimates and ANOVA results for the STB model on accuracy during training.

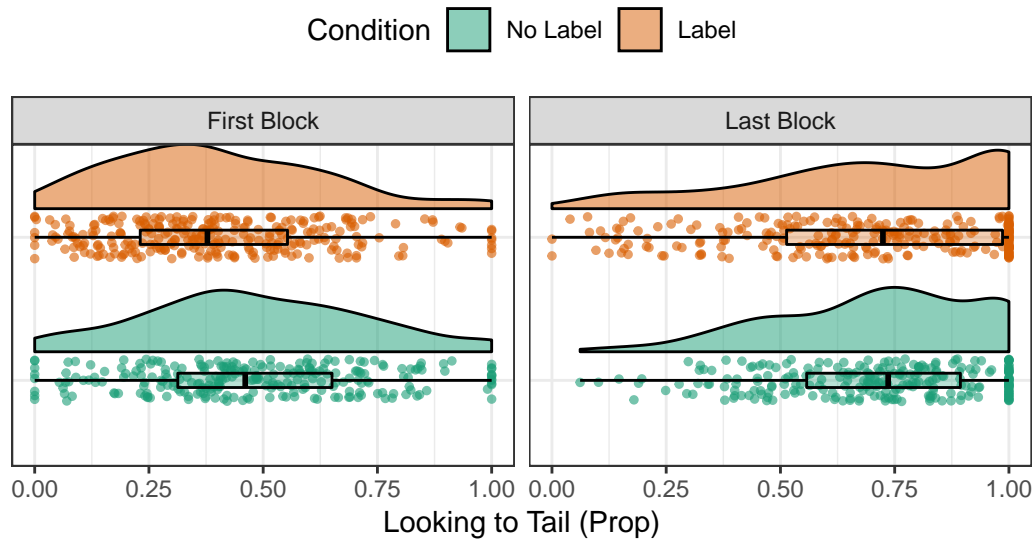


Figure 12: Raincloud plot from the data of the proportion of looking at the tail during an entire trial.

reaction time effects.

3.2.2 Eye-tracking Results

Proportion of Tail Looks by Trial We submitted proportion of looking to the tail during training to a linear mixed-effects regression model. The model included main effects of and interaction between Condition (no-label, label) and FstLst (first block, last block). The model also included random intercepts and slopes for FstLst by participant, and random intercepts by visual stimulus and by auditory stimulus. A summary of the parameter estimates and results of the ANOVA analysis on this model can be found in Table 7. A raincloud plot of the data can be seen in Fig. 12.

Only the main effect of FstLst reached significance, with participants in the no-label condition looking more towards the tail during the last block compared to the first block. There was no significant difference between the two groups, either during the first block or the last block.

| Parameter | Model Output | | ANOVA Output | |
|------------------|--------------|------------|--------------|-----------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 0.77 | 0.04 | | |
| FstLstLast Block | 0.30 | 0.05 | 86.91 | < .001 |
| ConditionLabel | -0.09 | 0.06 | 1.10 | 0.30 |
| FstLst:Condition | 0.07 | 0.07 | 0.85 | 0.36 |

Table 7: Parameter estimates and ANOVA results for the STB model on proportion looking at the tail after label onset.

| Parameter | Model Output | | ANOVA Output | |
|-----------------------------|--------------|------------|--------------|-----------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 0.78 | 0.04 | | |
| FstLstLast Block | 0.61 | 0.05 | 96.97 | < .001 |
| CurrentObjectFeedback | -0.02 | 0.05 | 44.12 | < .001 |
| CurrentObjectLabel | -0.08 | 0.06 | | |
| ConditionLabel | -0.08 | 0.09 | 1.29 | 0.28 |
| FstLst:COFeedback | -0.38 | 0.06 | 34.27 | < .001 |
| FstLst:COLabel | -0.27 | 0.08 | | |
| FstLst:Condition | 0.03 | 0.11 | 0.22 | 0.64 |
| COFeedback:Condition | -0.03 | 0.07 | 0.27 | 0.76 |
| COLabel:Condition | -0.01 | 0.08 | | |
| FstLst:COFeedback:Condition | 0.06 | 0.09 | 0.67 | 0.51 |
| FstLst:COLabel:Condition | -0.05 | 0.11 | | |

Table 8: Parameter estimates and ANOVA results for the STB model on proportion looking at the tail by trial window.

Proportion of Tail Looks by Trial Window To gain insight into how the feedback (non-linguistic in both groups and linguistic in the label group) influenced looking behaviour, we submitted proportion of looking at the tail during those different time windows for each trial to a linear mixed-effects regression model. The model included all main effects of and interactions between FstLst (first block, last block), CurrentObject (visual stimulus, feedback, label), and Condition (no-label, label). The model also included random intercepts by participant, visual stimulus (Stimulus), and category (StimLabel), and additional slopes by participant for FstLst, CurrentObject, and their interaction. A summary of the model’s parameter estimates and ANOVA analysis for those parameters can be seen in Table 8. Notably, the model gives us a parameter for each level of CurrentObject (and interactions including this effect), but the ANOVA analysis only computes an F value for this effect in general. A raincloud plot of the data can be seen in Fig. 13.

Participants in the no-label condition exhibited (a) more looking at the tail before categorisation in the last block, (b) less looking at the tail after categorisation (i.e. while hearing the feedback and after) during the first block, and (c) substantially less looking at the tail after categorisation during the last block (significant main effects of and interaction between FstLst and CurrentObject). No effect including Condition reached significance; in other words, participants in the label condition did not significantly differ from participants in the no-label condition at any point.

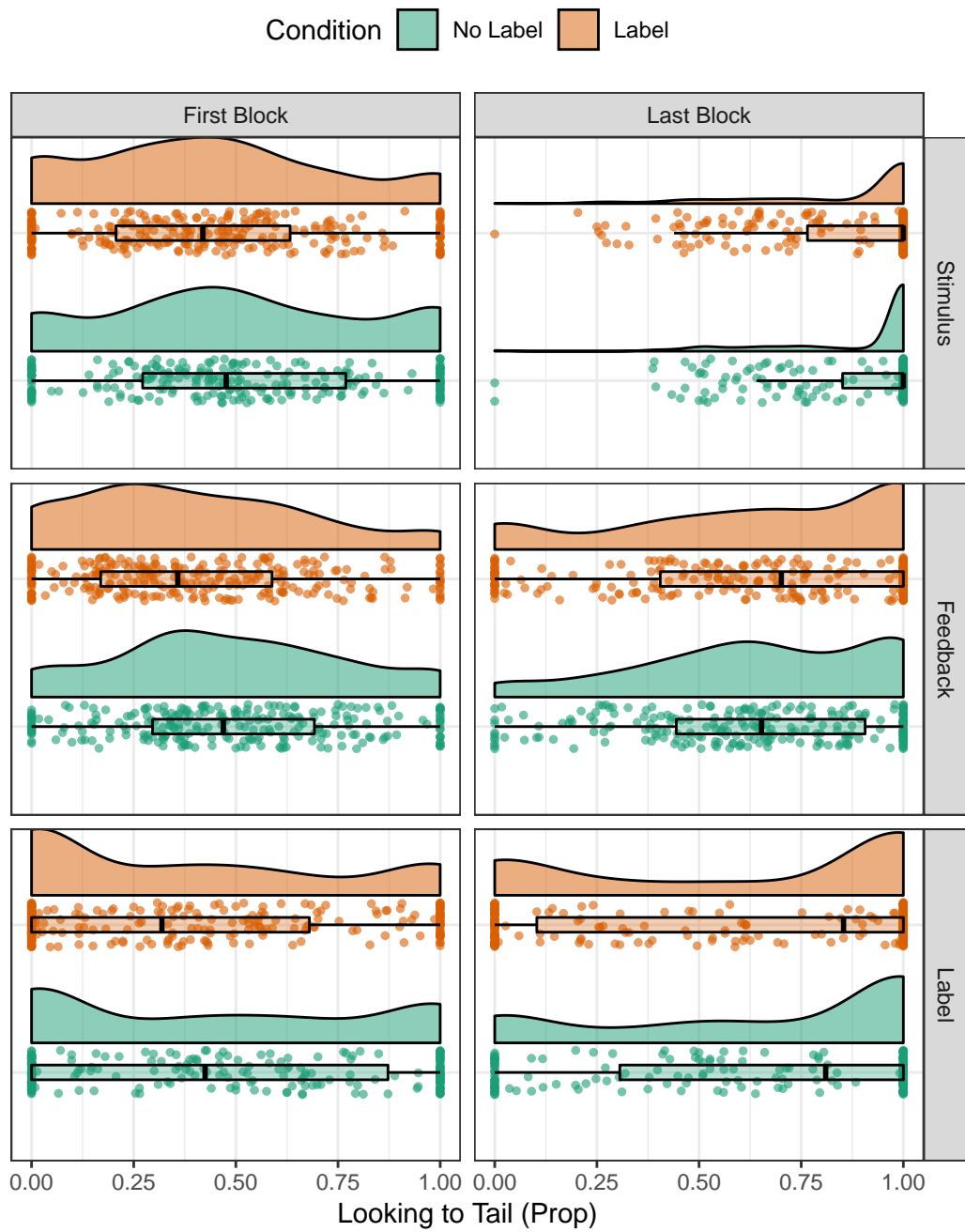


Figure 13: Raincloud plot from the data of the proportion of looking at the tail by trial window.

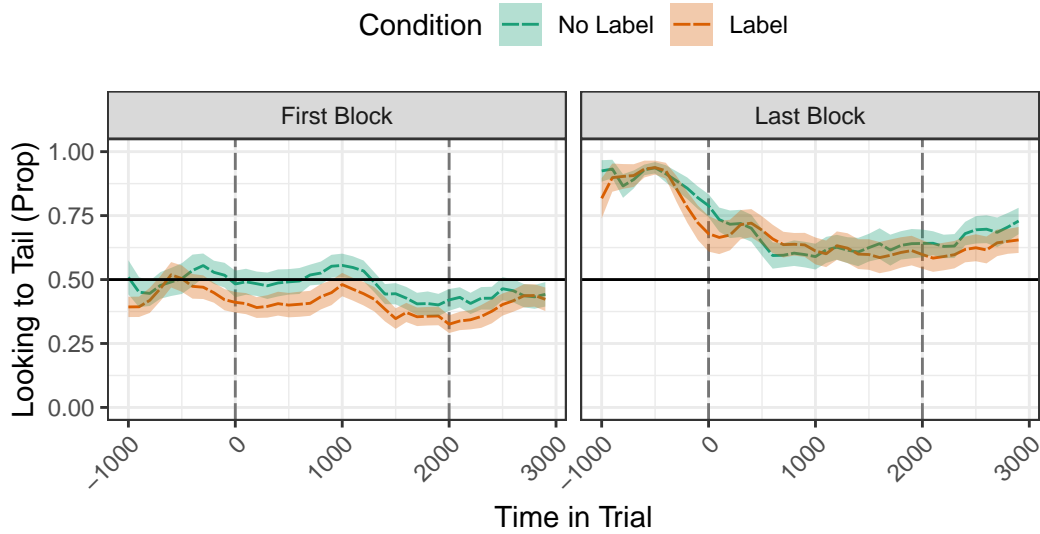


Figure 14: Time-course plot of the mean and SE of proportion looking at the tail. Vertical dashed line represent feedback onset (0 ms) and label onset (2000 ms).

Time-course Analysis Next, we analysed the evolution of proportion looking to the tail during training trials. We chose to include in the analysis only data from 1000 ms before button press until the end of the trial, as more than half of the reaction times were under 1000 ms (58.7%). A bootstrapped cluster-based permutation analysis revealed no difference between conditions either in the first or last block (see Fig. 14).

3.3 Discussion

In Experiment 2, we aimed to understand how auditory labels impact categorisation in adults, and whether or not this effect was different from what was observed in infants.

First, the behavioural results indicate that the task was too easy for adults, leading to a ceiling effect in terms of categorisation learning and accuracy, thus rendering any effect of labelling undetectable. Although it is typical for such experimental designs to find similar accuracy levels at the end of the experiment between subjects, participants who hear a label after the non-linguistic feedback usually display better accuracy and lower reaction times earlier (e.g. Lupyan et al., 2007). In previous studies however, categories were more ambiguous, and not defined by a sole feature; our simpler category structure could explain why participants in our study were at ceiling. Creating categories defined by more than one feature in a probabilistic manner and/or making the category boundaries more ambiguous, while keeping the same overall structure with a salient-non-diagnostic feature, could mitigate the ceiling effect encountered in this study.

Despite this null result, we had expected some effect of labelling on looking behaviour. Once again we did not find this, and it was most likely due to the simplicity of the task; specifically, participants did not have to rely on the label to encode the relevant information about the stimuli and categories. This explanation is further supported by one of the participants in the label group spontaneously sharing that they quickly

understood that the tail was the diagnostic feature and quickly made the categorisation decision on this basis.

4 General discussion

In this paper, we set out to study the effect of auditory labels on categorisation both in 15-month-old infants and in adults. More specifically, we tested the hypothesis that labels can act as category markers by directing attention to diagnostic features for a category. We found no evidence of an effect of labelling on looking behaviour during categorisation in either infants or adults, and no evidence for behavioural differences in adults. Although the lack of evidence in adults can be explained by a ceiling effect due to the simplicity of the task, the question remains as to why we found no significant differences in infants when numerous previous studies found varying effects under different conditions. It is of course entirely possible that we simply lacked the statistical power to detect an existing difference, nonetheless it is interesting to consider what a true null effect would mean on a theoretical point of view.

On the one hand, in line with the *labels-as-symbols* theory, we would have expected infants in the label condition to learn to look more at the diagnostic tail and less at the salient head throughout the experiment. This was not the case: there was no significant difference in proportion of looking at the tail depending on whether or not infants heard a label. Two recent studies found contradicting evidence (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015a): in these studies, infants looked longer at the diagnostic feature of two-featured objects when hearing a label. Two key differences between these studies and our design could explain those conflicting results. First, in Althaus's work, objects were not separated into two categories, and the diagnosticity of a feature was defined as a low between-exemplar variability for this feature, with the idea that diagnostic features are features common to all exemplar of a given category, therefore reducing overall variability between exemplars. In our study, the diagnostic tails were no less variable than the non-diagnostic heads, and were diagnostic only in the labelled categorisation context. This difference in the definition of diagnosticity could explain why we obtained different results, since we measured different concepts. More importantly, in our experiment, we wanted to make sure that the salience difference would be strong enough and shared by all participants, and therefore made the choice to use a very salient head. This high salience might have prevented any label-induced shifts in looking behaviour in the competition for attentional control. In contrast, stimuli in Althaus's work were specifically designed to reduce any difference in salience between the two features. It would be interesting in future studies to use varying degrees of salience between features.

On the other hand, the *labels-as-features* theory predicts that labels should not have an impact on attentional focus, which is consistent with our findings. However, based on this null result we do not claim that we have evidence in favour of the *labels-as-features* theory. Indeed, the *labels-as-symbols* theory could be correct, but in the current study other factors may have had a stronger influence on attentional focus. Further

work is needed to ascertain whether or not labels can impact attentional focus in some conditions, in particular in more realistic settings where the salience of diagnostic features is not substantially lower than the salience of non-diagnostic features, as was the case in our experiment.

In addition to these theoretical concerns regarding the effect of auditory labels on categorisation, our findings provide some insight into the use of eye-tracking data for measuring learning mechanisms. Eye-tracking is particularly useful when it comes to studying young infants who cannot give clear behavioural responses, and eye-tracking data have been commonly used as a proxy for their attentional processes. As an extension to the attentional focus evidenced by eye-tracking data, it has been assumed that infants look longer at stimuli that require more encoding, thus more attention (Houston-Price & Nakai, 2004); hence, an implicit link was made between looking times and learning mechanisms. From our work however, it is unclear that infants would necessarily look longer at a stimuli that requires more encoding. Particularly, we know that infants encoded enough information about the tails to link the correct one to its corresponding name; yet in a novelty preference trial they showed no preference towards a tail that they had never seen before, and thus needed more effort to encode. Instead, it seems that infants could encode enough information without it affecting their overall looking behaviour. This is even more clear in adults who have higher encoding capacities. Overall these results suggest that eye-tracking data are not a pure reflection of learning processes, but of something more.

In light of our findings, future studies should focus on understanding more precisely how labels impact categorisation and how they might compete with other factors for attention control, if they can have an impact at all on attentional focus. In addition to this, further work should seek to deepen our understanding of what processes impact eye movements in both adults and infants, to improve how we understand eye-tracking results and design eye-tracking experiments.

References

- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63. <https://doi.org/10.12688/wellcomeopenres.15191.1>
- Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PloS one*, 9(7), e99670. <https://doi.org/10.1371/journal.pone.0099670>
- Althaus, N., & Plunkett, K. (2015a). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, 1–11. <https://doi.org/10.1111/desc.12358>
- Althaus, N., & Plunkett, K. (2015b). Timing matters: The impact of label synchrony on infant categorisation. *Cognition*, 139, 1–9. <https://doi.org/10.1016/j.cognition.2015.02.004>

- Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, *151*, 5–17. <https://doi.org/10.1016/j.jecp.2015.11.013>
- Aslin, R. N. (2007). What’s in a look? *Developmental Science*, *10*(1), 48–53. <https://doi.org/10.1111/j.1467-7687.2007.00563.x>
- Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Effects of linguistic labels on visual attention in children and young adults. *Frontiers in Psychology*, *9*. <https://doi.org/10.3389/fpsyg.2018.00358>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bronson, G. W. (1991). Infant differences in rate of visual encoding. *Child Development*, *62*(1), 44. <https://doi.org/10.2307/1130703>
- Colombo, J., Mitchell, D. W., Coldren, J. T., & Freeseaman, L. J. (1991). Individual differences in infant visual attention: Are short lookers faster processors or feature processors? *Child Development*, *62*(6), 1247. <https://doi.org/10.2307/1130804>
- Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, *23*(2), 178–186. <https://doi.org/10.1177/0956797611429133>
- Deng, W., & Sloutsky, V. M. (2015). Linguistic labels, dynamic visual features, and attention in infant category learning. *Journal of Experimental Child Psychology*, *134*, 62–77. <https://doi.org/10.1016/j.jecp.2015.01.012>
- Dink, J., & Ferguson, B. (2018). *eyetrackingR* [R package version 0.1.8]. R package version 0.1.8.
- Gelman, S. A., & Coley, J. D. (1991). Language and categorization: The acquisition of natural kind terms. In *Perspectives on language and thought: Interrelations in development* (pp. 146–196). Cambridge, Cambridge University Press.
- Gliga, T., Volcin, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, *22*(12), 2781–2789. <https://doi.org/10.1162/jocn.2010.21427>
- Graham, S. A., & Poulin-Dubois, D. (1999). Infants’ reliance on shape to generalize novel labels to animate and inanimate objects. *Journal of child language*, *26*(2), 295–320. <https://doi.org/10.1017/S0305000999003815>
- Hilton, M., Twomey, K. E., & Westermann, G. (2019). Taking their eye off the ball: How shyness affects children’s attention during word learning. *Journal of Experimental Child Psychology*, *183*, 134–145. <https://doi.org/10.1016/j.jecp.2019.01.023>
- Hilton, M., & Westermann, G. (2017). The effect of shyness on children’s formation and retention of novel word–object mappings. *Journal of Child Language*, *44*(6), 1394–1412. <https://doi.org/10.1017/S030500091600057X>

- Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. <https://doi.org/10.1002/icd.364>
- Jankowski, J. J., Rose, S. A., & Feldman, J. F. (2001). Modifying the distribution of attention in infants. *Child Development*, *72*(2), 339–351. <https://doi.org/10.1111/1467-8624.00282>
- Kovic, V., Plunkett, K., & Westermann, G. (2009). Eye-tracking study of animate objects. *Psihologija*, *42*(3), 307–327. <https://doi.org/10.2298/PSI0903307K>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
- Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means* [R package version 1.3.5.1]. R package version 1.3.5.1.
- Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & Frank, M. (2016). A quantitative synthesis of early language acquisition using meta-analysis. *PsyArXiv*. <https://doi.org/10.17605/osf.io/htsjm>
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. <https://doi.org/10.1111/j.1467-9280.2007.02028.x>
- Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06
- Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645. <https://doi.org/10.1037//0012-1649.36.5.635>
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681. <https://doi.org/10.1016/j.cognition.2007.04.003>
- Quinn, P. C., Doran, M. M., Reiss, J. E., & Hoffman, J. E. (2009). Time course of visual attention in infant categorization of cats versus dogs: Evidence for a head bias as revealed through eye tracking. *Child Development*, *80*(1), 151–161. <https://doi.org/10.1111/j.1467-8624.2008.01251.x>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2003). Infant visual recognition memory: Independent contributions of speed and attention. *Developmental Psychology*, *39*(3), 563–571. <https://doi.org/10.1037/0012-1649.39.3.563>
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166–188. <https://doi.org/10.1037/0096-3445.133.2.166>

- Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child development*, 1695–1709. <https://doi.org/10.1111/1467-8624.00373>
- Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, 42, 413–438. <https://doi.org/10.1111/cogs.12539>
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634), 20120391–20120391. <https://doi.org/10.1098/rstb.2012.0391>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Chapter 4

Labels Drive Adults' Attention to Salient Features

In the previous chapter, we used the same experimental design on 15-month-old infants and adults to study the potential effects on visual attention of feature salience and object labelling during a categorisation task. We presented participants with categories in which the salient feature (the head of line-drawn novel animals) was non-diagnostic of category membership, but the non-salient feature (tail) was diagnostic. We found that participants who heard a label (redundant for adults, as they were also given non-verbal feedback on their categorisation choice) could learn those counter-intuitive categories without looking differently at the stimuli from participants in a control group who did not hear a label.

In infants, this result was interpreted as indirect evidence against the *labels-as-symbols* theory and thus in favour of the *labels-as-features* theory. Most importantly, taken together with the rest of the literature, this finding hinted that eye movements and looking times are not a good proxy for learning, but only measure visual attention patterns.

In adults, however, there was clear evidence from their accuracy measure that participants were at ceiling performance, and this could at least partly explain the absence of differences in looking patterns. To further study this question in adults, we thus extended our experimental design to make it more complex for adults. Specifically, we increased the number of non-salient diagnostic features that participants needed to consider for category learning. We present this new design and the ensuing results in this chapter.

A Name for a Head: Auditory Labels Drive Adults' Attention to Salient, Not Diagnostic, Object Features

Arthur Capelier-Mourguy, Ho Yeung, Katherine E. Twomey, Gert Westermann
Department of Psychology, Lancaster University (UK)

Abstract

The effect of auditory labels on category formation has been studied extensively in infants and adults. Although there is a consensus that adults see labels as symbolic category markers, it is unclear whether infants process labels in such a way early in development, as argued by the *labels-as-symbols* theory, or if infants first see labels as object features, and slowly learn to give them a more symbolic value, as argued by the *labels-as-features* theory. An important prediction of the *labels-as-symbols* theory is that labels should highlight diagnostic features. This prediction was recently tested on adults and 15-month-old infants, but resulted in a ceiling effect in adults. Here, we extend this previous study to mitigate this ceiling effect. Precisely, we presented adults with animal drawing where the non salient feet and tail were diagnostic of category membership, but the salient head was not. We found that adults who heard category labels looked more at the non-diagnostic head. Considering that head are usually diagnostic of category membership in real-life, this suggests that the effect of auditory labels on attention during categorisation tasks is heavily influenced by background knowledge.

Keywords: labelling, categorisation, salience, diagnosticity

1 Introduction

A key component of human cognition is the ability to bring objects we encounter together into categories, to reduce the cognitive cost of processing new exemplars of those categories. Starting as early as 10-month-old, infants automatically group together items that are similar, and separate items that are dissimilar, slowly building up categories based on what they see (Mareschal & French, 2000; Mareschal et al., 2000). Category

exemplars are often encountered together with the name of the category spoken by a caregiver early in development, and such naming events also have been argued to improve categorisation in infants and adults (e.g. Althaus & Westermann, 2016; Balaban & Waxman, 1997; S. A. Gelman & Coley, 1991; Gliga et al., 2010; Graham & Poulin-Dubois, 1999; Lupyan et al., 2007; Plunkett et al., 2008; Waxman & Markow, 1995). However, the mechanism by which adding a spoken label improves categorisation processes early in development, and how those mechanisms develop, remains unclear. Two main theories attempt to explain the role of labels in categorisation early in development: the *labels-as-symbols* view argues that labels are category markers, while the *labels-as-features* view argues that labels do not differ from other object features. Both theories however agree that later in development adults see labels as category markers, and the *compound-representations* theory offers a mechanism to account for a developmental change from a featural to a symbolic role of labels.

On the one hand, the *labels-as-symbols* theory suggests that labels are, from an early developmental stage, abstract decontextualised cues that are separate from object representations and act as referential pointers in a top-down way, inviting the listener to form categories (Waxman & Markow, 1995). A possible mechanism for this theory is that labels drive attention towards diagnostic features, that is, features shared by all exemplars of the category but not by out-of-category items. For example, knowing that both humans and elves walk on two legs, use tools, and talk, is not helpful in discriminating them into two categories, while the body hair of humans and the pointy ears of elves are both diagnostic features for their respective categories. Supporting this theory, studies in infancy research have shown that adding a label specifically allowed infants to form categories that they would not otherwise form (Althaus & Westermann, 2016; Plunkett et al., 2008; Waxman & Markow, 1995), and increased infants' attention towards and encoding of diagnostic features (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015). In adults, studies have shown that participants attended selectively to diagnostic features when hearing a label (Best et al., 2013), that auditory labels reliably primed category representations across participants whereas other non-linguistic sounds only primed specific exemplars of those categories with between-subject differences in the exemplars primed (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012), and finally that the enhancement of labelling by up-regulating activity over Wernicke's

area via transcranial direct current stimulation selectively improved the formation of “sparse” categories heavily relying on a few diagnostic features, and vice versa when disturbing labelling (Perry & Lupyan, 2014, 2016).

On the other hand, the *labels-as-features* theory suggests that labels are first treated as features, part of the object representation at the same level as other physical or auditory features: a dog is an animal with four legs, fur, a tail, a dog face, and is called “dog” (Sloutsky & Fisher, 2004). In this theory, labels simply facilitate categorisation by adding to the overall similarity of all exemplars within a category, since they all share the same name in addition to other features. However, the *labels-as-features* theory mostly applies to the earlier stages of development, and proponents of this account agree that label perception evolves to bear a more symbolic role later in development (e.g. Best et al., 2013; Deng & Sloutsky, 2016; Sloutsky & Fisher, 2012). In particular, a study with different age groups noted that 4- to 5-year-old children perceived labels as features, 11- to 12-year-old children perceived labels as symbolic markers, and 7- to 8-year-old children were in a transitional stage with some children being more feature respondent and others more symbolic marker respondent (Sloutsky et al., 2001). Nevertheless, there is some evidence that adults can still treat labels as features in some contexts, suggesting that the mechanisms by which labels are perceived as symbols does not necessarily replace the initial role of labels as features (Deng & Sloutsky, 2012).

A third view introduced recently can account for this evolution in time (Westermann & Mareschal, 2014). This *compound-representations* account assumes that labels are encoded in the same representational space as other features, but are not integrated into the object representations, only linked to them. In this way, labels will first drive categorisation by adding to the within-category similarity. With learning, over time, labels will become more closely associated to object representations, and act more like markers for categories, reducing the distance in representational space between exemplars of the same category.

Despite the numerous studies conducted on this topic in adults, no study has focused specifically on the online process of category learning and effects of labels on this process, to our knowledge. Many studies have focused on behavioural measures such as categorisation accuracy, reaction time, or inference of a missing feature (e.g. Best et al., 2013; Deng & Sloutsky, 2016; Edmiston & Lupyan, 2015; Lupyan et al., 2007;

Lupyan & Thompson-Schill, 2012; Sloutsky & Fisher, 2012). Other studies have used eye-tracking and found preferential looking to a target amongst a set of items when hearing a redundant, task-irrelevant label (Edmiston & Lupyan, 2015; Lupyan, 2008; Lupyan & Spivey, 2010; Salverda & Altmann, 2011). Critically, the existing research with adults does not present a clear picture of how category labels affect the attention to and processing of object features of different salience when learning novel categories. A recent study with infants and adults addressed these questions by asking whether or not a label could actively guide categorisation in categories where low-salience, but not high-salience, features were diagnostic (Capelier-Mourguy et al., 2019). In this study, 15-month-old infants and adults were presented with the same stimuli: a series of simple two-featured snake-like animals, with a salient head and non-salient a tail. Importantly, the high-salience head did not indicate category membership but varied pseudorandomly during training; in contrast, the low-salience tail was diagnostic of category membership. Adults further had to make a categorisation choice for each exemplar, and were given non-linguistic auditory feedback. Based on previous literature, the authors hypothesised that adults hearing a redundant label would look more at the diagnostic tail and would be better and quicker at learning the categories than adults hearing only an auditory feedback after categorising each animal. These predictions were not upheld: using the same stimuli and category structure for both 15-month-old infants and adults led to a ceiling effect in adults in terms of learning speed and accuracy, and an absence of difference in terms of looking patterns.

In the current study we aimed to mitigate this ceiling effect in adults by extending Capelier-Mourguy et al.'s work (2019, hereafter CMTW). We did so by using two non-salient features instead of only one (a tail and feet), and making neither of them fully diagnostic, so that participants would have to pay attention to both features to learn categories. As in CMTW, we expected that (a) participants who heard a label would, during the training phase, look more and/or more quickly at the diagnostic features, and encode them more robustly, and (b) participants who heard a label would form categories more quickly.

2 Methods

All materials used for this experiment are available online for inspection and replication purposes¹, including stimuli, the experiment script in Eprime (version 2), raw data, and analysis scripts in R.

2.1 Data Handling and Software Specifications

Data Handling A common measure in eye-tracking data analysis is the proportion of looking at an area of interest (AOI). To account for the boundedness of proportion values, we used the arcsine-root transformation of the proportion in our statistical models; for ease of discussion, we use the term “proportion” to talk about this measure. However, we plot raw proportion values only, for ease of visual interpretation.

Further, we discarded looks outside of our defined AOIs. This means that, for example, the proportion of looking at the tail during the familiarisation trials is defined as the time spent looking at the tail divided by the time spent looking at the tail, feet, or head, but not the total time spent looking at the screen during a trial.

Software Specifications All statistical results were obtained using R (version 3.6.1; R Core Team, 2019). Analyses in this paper were conducted using (a) `lme4` (version 1.1-17; Bates et al., 2015) to run Sample Theory Based (STB) (generalised) linear mixed-effects models, `lmerTest` (version 3.0-1; Kuznetsova et al., 2017) to run ANOVA analyses on those mixed-effect models, and the `p.adjust` function from the base `stats` package to adjust p -values when needed, (b) `eyetrackingR` (version 0.1.8; Dink & Ferguson, 2018) to handle eye-tracking data and run bootstrapped cluster-based permutation analyses, and (c) `ggplot` (version 2.2.1; Wickham, 2016) to plot graphs from our data and `ggeffects` (version 2.4.1; Lüdtke, 2018) to compute and plot estimated marginal effects from our models.

2.2 Participants

We recruited 40 participants from Lancaster University via an online pool of participants for psychology studies. After exclusion of four outliers (more than two standard deviations above the mean) in terms of learning speed, the final sample for behavioural

¹<https://osf.io/5yh67/>

| Category | Tail | Feet | Head |
|----------|----------|----------|------|
| <i>A</i> | α | ν | 1 |
| | α | ν | 2 |
| | α | α | 1 |
| | α | α | 2 |
| | ν | α | 1 |
| | ν | α | 2 |
| <i>B</i> | β | ν | 1 |
| | β | ν | 2 |
| | β | β | 1 |
| | β | β | 2 |
| | ν | β | 1 |
| | ν | β | 2 |

Table 1: Category structure. α represents a feature belonging to category *A*, β a feature belonging to category *B*, and ν a neutral feature. Heads are never diagnostic and so the two different versions are coded 1 and 2

results consisted of 36 participants (17 female, $M_{\text{age}} = 20.58$, range 19-23). A further five participants did not contribute to eye-tracking data due to not meeting our inclusion criteria (minimum 70% of looking on 70% of the trials). All participants were fluent in English.

2.3 Materials

Visual Stimuli We used structurally similar stimuli as those in CMTW, adding a third low-salient feature.

Our stimuli thus consisted of simple snake-like animals with three features only: a salient head (Kovic et al., 2009), and two non-salient features, a tail and feet. This ensured that stimuli afforded a “natural” non-uniform salience shared by all participants. While the salient head was never diagnostic of category membership, we designed stimuli so that neither the feet nor tail were fully diagnostic. To do so, we defined two prototypes for the non-diagnostic head, but three prototypes for feet and tail: one prototype for each category, and one “neutral” prototype that would correspond to neither category. See Table 1 for a structural description of the stimuli, and Fig. 1 for examples of stimuli displaying all possible feature versions.

Auditory Stimuli After categorising each exemplar, participants in both conditions were given auditory feedback in the form of a shimmering sound for correct or a buzzer for incorrect categorisation. Then, participants in the label condition heard the phrase

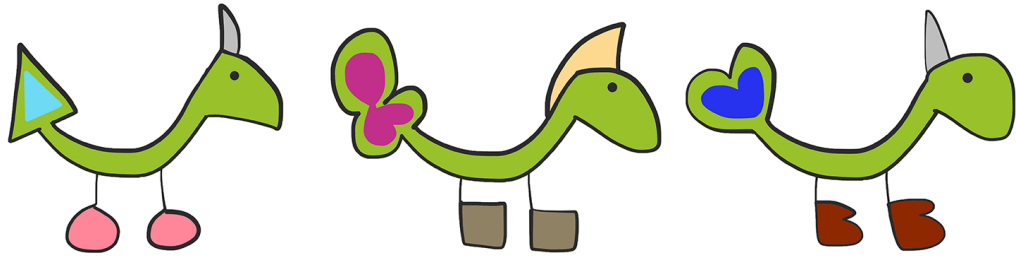


Figure 1: Example of stimuli displaying all three kinds of feet and tail, and both kinds of heads.

“It’s a [Saldie/Gatoo]”, pronounced by a female native British speaker in a neutral tone. Both feedback sounds lasted for 2000 ms, and both labelling phrases lasted for 1000 ms with a label onset at 400 ms.

2.4 Procedure and Design

Participants were tested in a quiet room, using a Tobii X120 eye-tracker calibrated using a 9-point routine to record eye-tracking data, and Eprime to run the experiment and collect behavioural data (categorisation responses, reaction time, number of training blocks, etc.).

The experiment consisted of a categorisation task: participants were presented with one exemplar at a time, and were asked to sort them into one of two categories by pressing the corresponding button on a keyboard. Participants were first presented with a training phase, during which they were provided with feedback after each categorisation decision. Participants in all groups heard non-linguistic feedback, followed by the label for the category for half of the participants. This training phase lasted for up to 21 blocks, or until successful categorisation (i.e. one full block without any mistakes). Each block consisted of the 12 exemplars described in Table 1 presented in a random order. A fixation cross was presented in the middle of the screen for one second before each trial.

Participants were then presented with a test phase that consisted of the same categorisation task for one block without feedback.

3 Results

Analysis Structure We conducted two types of analysis in this report: testing average proportion looking during one or several time windows of a trial, and time-course analysis. We also tested for other unique-per-trial values, however these tests followed the same structure as tests on proportion looking.

For the tests of proportion looking, we used (generalised) linear mixed-effects regression models fitted with maximal converging random-effects structure to estimate parameters (Barr et al., 2013). For significance testing of those parameters, we used type I ANOVA analyses with Satterthwaite’s method as implemented in `lmerTest` for linear models, and commonly-used asymptotic Wald tests for generalised linear models.

For the time-course analyses, we used bootstrapped cluster-based permutation analysis as implemented in `eyetrackingR` with 100 ms time bins and t -test comparisons between the two conditions (no-label, label); the choice of a t -test rather than a mixed-effects model was due to the current implementation in `eyetrackingR` that did not allow for the use of mixed-effects models when testing a between-subject factor as in our case. To test different levels of other factors (e.g. first three trials against last three trials), we ran an independent analysis on each level of this factor (or levels of their interaction when using multiple factors); although this approach involved multiple comparisons, there is to our knowledge no straightforward way to test for multiple factors directly.

In the same way that conducting multiple independent t -tests after a significant ANOVA interaction increases the likelihood of a type I error, testing for a great number of parameters in a single model (e.g. regression or ANOVA) increases the chance of finding significant p -values by chance (see Shaffer, 1995, for a review on multiple hypothesis testing). However, if the use of corrected p -values, or q -values, is consensual for multiple post-hoc tests (e.g. the Bonferroni adjustment introduced by Dunn, 1961), the question of when and how to correct for multiple tests for regression parameters is still debated (e.g. A. Gelman et al., 2012). Considering that we are here often testing for a great number of parameters, we provide uncorrected p -values for our models, but contrast them with q -values based on the less stringent control of false discovery rate (thereafter ‘fdr’) proposed by Benjamini and Hochberg (1995) rather than more conservative family-wise error rate adjustment methods such as the Bonferroni correction. To keep this

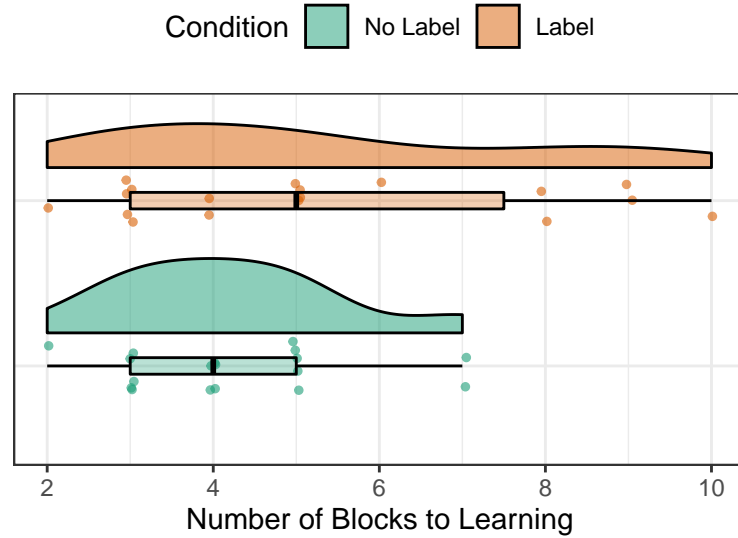


Figure 2: Raincloud plot from the data of the number of blocks to learning.

article clearer, we only use fdr q -values for p -values that are only mildly significant; the uncorrected p -values further allow readers to apply the adjustment method they think the fittest, and thus change their interpretation of our results accordingly.

3.1 Behavioural Results

One of our hypotheses was that participants' categorisation abilities would benefit from hearing a label on top of the 'correct/wrong' auditory feedback. This would be reflected in the number of training blocks they need to learn the categories (i.e. complete a full block without any categorisation mistakes), as well as in the overall response accuracy throughout training.

Number of Blocks to Learning Given that this measure was not normally distributed (Anderson-Darling normality test: $A = 1.6971$, $p = .0002$), we first conducted an independent 2-group Mann-Whitney U-test to test the effect of label on the number of training blocks to learning. We found no differences between the label and no-label group ($W = 129$, $p = .2929$), and as such, there was no evidence that labels helped participants learn categories more quickly. A "raincloud" plot (Allen et al., 2019) of the data is shown in Fig. 2. These plots include half a violin plot to understand the shape of the data, individual data points to better understand the structure of the data, and a boxplot to give some descriptive statistics at a glance.

| Parameter | Estimate | Std. Error | z value | $Pr(> z)$ |
|---------------------------------------|----------|------------|-----------|-------------|
| (Intercept) | 1.56 | 0.50 | 3.11 | 0.002 |
| Block | 1.46 | 0.42 | 3.50 | .001 |
| zLogRT | -0.33 | 0.40 | -0.82 | 0.411 |
| DiagnosticFeet | -1.41 | 0.55 | -2.55 | 0.011 |
| DiagnosticTail | -0.51 | 0.54 | -0.95 | 0.344 |
| Condition | -0.31 | 0.62 | -0.49 | 0.622 |
| Block:zLogRT | -0.67 | 0.41 | -1.64 | 0.100 |
| Block:DiagnosticFeet | -0.51 | 0.44 | -1.16 | 0.246 |
| Block:DiagnosticTail | -1.00 | 0.45 | -2.23 | 0.026 |
| zLogRT:DiagnosticFeet | 0.35 | 0.47 | 0.75 | 0.455 |
| zLogRT:DiagnosticTail | 0.17 | 0.49 | 0.35 | 0.725 |
| Block:Condition | -0.76 | 0.41 | -1.85 | 0.065 |
| zLogRT:Condition | -0.15 | 0.53 | -0.29 | 0.775 |
| DiagnosticFeet:Condition | 0.33 | 0.65 | 0.50 | 0.614 |
| DiagnosticTail:Condition | -0.19 | 0.63 | -0.30 | 0.765 |
| Block:zLogRT:DiagnosticFeet | 0.37 | 0.43 | 0.86 | 0.389 |
| Block:zLogRT:DiagnosticTail | 0.56 | 0.44 | 1.28 | 0.202 |
| Block:zLogRT:Condition | 0.31 | 0.42 | 0.74 | 0.460 |
| Block:DiagnosticFeet:Condition | 0.27 | 0.43 | 0.63 | 0.531 |
| Block:DiagnosticTail:Condition | 0.69 | 0.44 | 1.55 | 0.120 |
| zLogRT:DiagnosticFeet:Condition | 0.06 | 0.65 | 0.10 | 0.922 |
| zLogRT:DiagnosticTail:Condition | -0.01 | 0.66 | -0.01 | 0.992 |
| Block:zLogRT:DiagnosticFeet:Condition | -0.15 | 0.44 | -0.33 | 0.740 |
| Block:zLogRT:DiagnosticTail:Condition | -0.32 | 0.45 | -0.70 | 0.486 |

Table 2: Summary of the glmer model for accuracy during training.

Accuracy We submitted response accuracy to a binomial generalised linear mixed-effects restricted model. The model included all main effects of and interactions between Block (numeric), Reaction Time (zLogRT, log-transformed and scaled), Diagnostic feature (both, feet, tail), and Condition (no-label, label). The model also included random intercept and slopes for Block, zLogRT, Diagnostic, and their interactions, by participant; and random intercept by visual stimulus and by auditory stimulus. The parameter estimates for this model are given in Table 2.

The most notable significant effect here is that of Block, with accuracy increasing throughout training for participants in the no-label condition when they had average reaction times when categorising exemplars where both the feet and tail were diagnostic. Other than that, two p -values reached significance: for the main effect of DiagnosticFeet, with lower accuracy when only the feet were diagnostic during the first block in the no-label condition for average reaction times compared to when both features were diagnostic, and for the Block-by-DiagnosticTail interaction with a slower increase in ac-

curacy in the no-label group for average looking times compared to when both features were diagnostic. However, the *fdr* correction gave non-significant *q*-values for the two effects considered here ($q = 0.087$ and $q = 0.154$ respectively).

Notably, none of the other parameters were significant. Thus, accuracy did not significantly differ with respect to any other variable or interactions, and in particular, participants in the label condition did not significantly differ from participants in the no-label condition at any point regardless of what features were diagnostic and regardless of their reaction times. In conclusion, participants first had more difficulties successfully categorising exemplars for which only the feet were diagnostic, but by the end of training participants successfully categorised all exemplars, all that regardless of the presence or absence of auditory labels. Put differently, labels did not help participants reach higher accuracy earlier in training.

3.2 Eye-tracking Results

Average Proportion of Looking We submitted proportion of looking to the different AOIs during training to a linear mixed-effects model. The model included main effects of and interaction between *FstLst* (first block, last block), *AOI* (head, feet, tail), *Diagnostic* feature (both, feet, tail), and *Condition* (no-label, label). The model also included random intercepts and slopes for *FstLst*, *AOI*, *Diagnostic*, and their interactions, by participant; and random intercepts by visual stimulus and by auditory stimulus. A summary of the parameter estimates and results of the ANOVA analysis on this model can be found in Table 3. Note that while some of our variables were categorical with multiple levels (*AOI*, *Diagnostic* and associated interactions), the ANOVA analysis only computed an *F* value for these effects as a whole. We therefore report the *F* value and associated *p*-value on the first line only for categorical variables and associated interactions with more than two levels. A raincloud plot of the data can be seen in Fig. 3.

First, we saw a significant main effect of *AOI*, with participants looking much less at the feet and more at the tail compared to the head in the no-label condition during the first block of training when both features were diagnostic. Looks towards the two diagnostic features increased by the end of training for those same participants when both features were diagnostic, as evidenced by the significant *FstLst*-by-*AOI* interac-

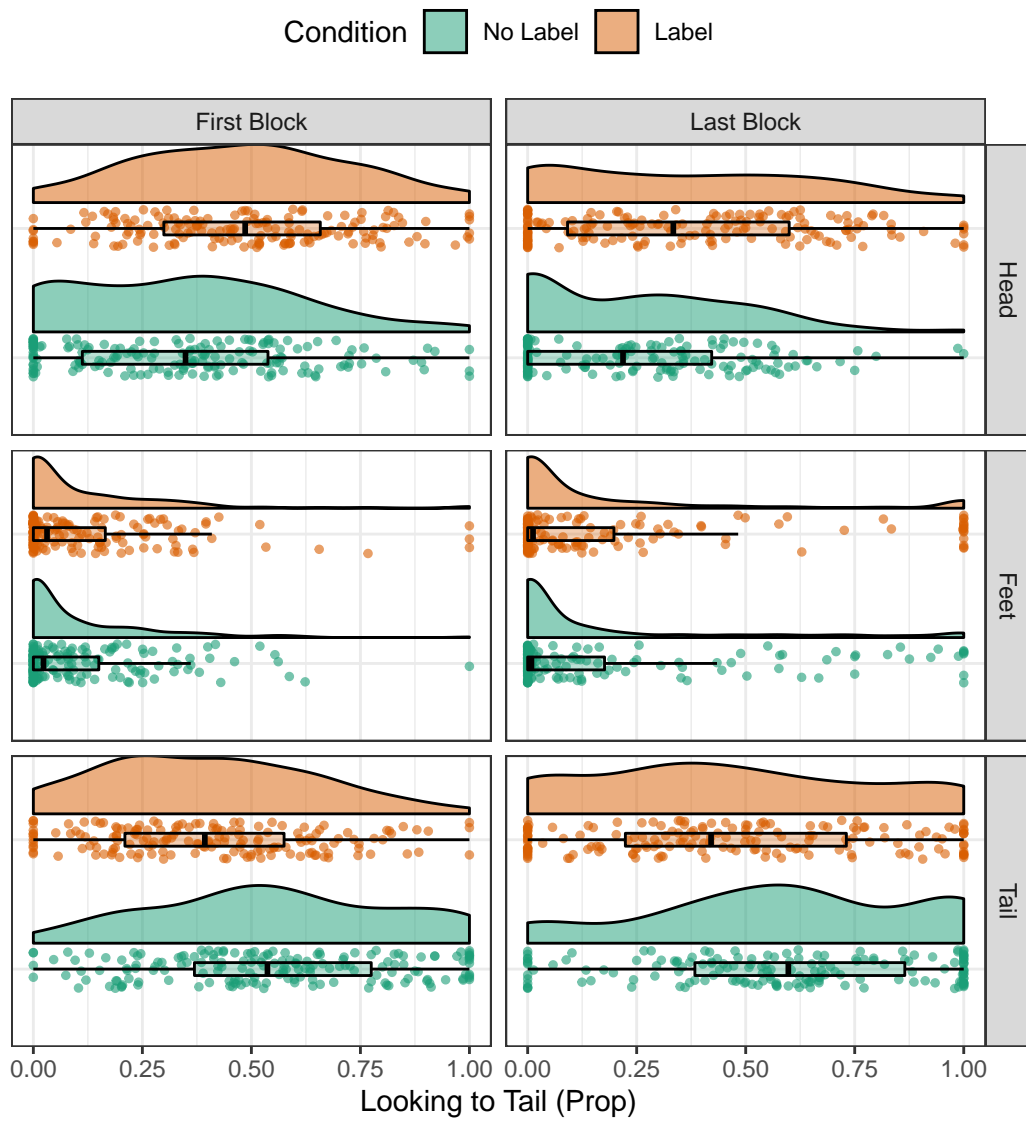


Figure 3: Raincloud plot from the data of the proportion of looking at the different AOIs.

| Parameter | Model Output | | ANOVA Output | |
|---|--------------|------------|----------------|-------------------------|
| | Estimate | Std. Error | <i>F</i> value | <i>Pr</i> (> <i>F</i>) |
| (Intercept) | 0.59 | 0.07 | | |
| FstLst | -0.13 | 0.09 | 0.98 | 0.323 |
| AOIFeet | -0.38 | 0.10 | 25.33 | .001 |
| AOITail | 0.28 | 0.11 | | |
| DiagnosticFeet | 0.02 | 0.06 | 0.02 | 0.979 |
| DiagnosticTail | -0.05 | 0.06 | | |
| Condition | 0.13 | 0.09 | 0.22 | 0.642 |
| FstLst:AOIFeet | 0.17 | 0.12 | 6.06 | 0.006 |
| FstLst:AOITail | 0.18 | 0.17 | | |
| FstLst:DiagnosticFeet | -0.02 | 0.10 | 0.02 | 0.984 |
| FstLst:DiagnosticTail | -0.01 | 0.09 | | |
| AOIFeet:DiagnosticFeet | 0.05 | 0.09 | 1.51 | 0.224 |
| AOITail:DiagnosticFeet | -0.08 | 0.09 | | |
| AOIFeet:DiagnosticTail | 0.06 | 0.08 | | |
| AOITail:DiagnosticTail | 0.11 | 0.08 | | |
| FstLst:Condition | -0.01 | 0.12 | 0.01 | 0.908 |
| AOIFeet:Condition | -0.05 | 0.14 | 4.23 | 0.024 |
| AOITail:Condition | -0.33 | 0.16 | | |
| DiagnosticFeet:Condition | 0.04 | 0.09 | 0.06 | 0.939 |
| DiagnosticTail:Condition | 0.11 | 0.08 | | |
| FstLst:AOIFeet:DiagnosticFeet | 0.04 | 0.14 | 0.69 | 0.605 |
| FstLst:AOITail:DiagnosticFeet | 0.00 | 0.17 | | |
| FstLst:AOIFeet:DiagnosticTail | 0.11 | 0.13 | | |
| FstLst:AOITail:DiagnosticTail | -0.06 | 0.15 | | |
| FstLst:AOIFeet:Condition | -0.03 | 0.17 | 0.03 | 0.975 |
| FstLst:AOITail:Condition | 0.06 | 0.24 | | |
| FstLst:DiagnosticFeet:Condition | 0.00 | 0.14 | 0.06 | 0.946 |
| FstLst:DiagnosticTail:Condition | 0.02 | 0.13 | | |
| AOIFeet:DiagnosticFeet:Condition | -0.16 | 0.12 | 1.53 | 0.219 |
| AOITail:DiagnosticFeet:Condition | -0.01 | 0.13 | | |
| AOIFeet:DiagnosticTail:Condition | -0.17 | 0.11 | | |
| AOITail:DiagnosticTail:Condition | -0.17 | 0.12 | | |
| FstLst:AOIFeet:DiagnosticFeet:Condition | 0.10 | 0.20 | 0.60 | 0.667 |
| FstLst:AOITail:DiagnosticFeet:Condition | -0.09 | 0.24 | | |
| FstLst:AOIFeet:DiagnosticTail:Condition | -0.09 | 0.19 | | |
| FstLst:AOITail:DiagnosticTail:Condition | -0.01 | 0.20 | | |

Table 3: Summary of the lmer model for proportion looking at the different AOIs during training.

| AOI | Direction | Cluster Position | Summed Statistic | Probability |
|------|------------------|------------------|------------------|-------------|
| Head | label > no-label | 600 - 1200 ms | -14.39 | 0.048 |
| | label > no-label | 1800 - 2400 ms | -19.32 | 0.026 |
| Tail | no-label > label | 1800 - 2600 ms | 27.64 | 0.016 |

Table 4: Summary of the bootstrapped cluster-based permutation analysis on proportion of looking at the different AOIs.

tion. Furthermore, there was a mildly significant AOI-by-Condition interaction, with participants in the label condition looking less at the tail compared to participants in the no-label condition in the first block of training when both features were diagnostic. However, the *fdr* correction gave a non-significant $q = 0.122$.

No other effects were significant. Notably, looking patterns did not differ depending on which features were diagnostic at any time for participants in the no-label condition, and the difference in looking pattern between participants in the label and no-label condition did not differ significantly depending on which features were diagnostic or between the first and last training block. In other words, participants overall looked much less at the feet than other AOIs, but looked more at both diagnostic features (tail and feet) by the end of training, and participants in the label condition might have been looking less at the tail compared to participants in the no-label condition.

Time-course Analysis To understand better how participants divided their attention between the three AOIs during training, we ran one bootstrapped cluster-based permutation analysis for each AOI for the first and last block of training. We chose to include in the analysis only data from 1000 ms before button press until the end of the trial, as more than half of the reaction times were under 1000 ms (50.4%). The clusters that reached significance are displayed in Fig. 4, and *p*-values for those clusters are reported in Table 4.

From this analysis, we can see that participants in the label condition looked more at the head and less at the tail in the first block of training compared to participants in the no-label condition, and participants in both conditions looked equally little at the feet. There seemed to be an overall similar trend in the last block of training, however with no clusters reaching significance. We return to the temporal location of the significant clusters in the first block in the general discussion, it does seem however that differences arose after categorisation, and that the presence of auditory labels elicited increased

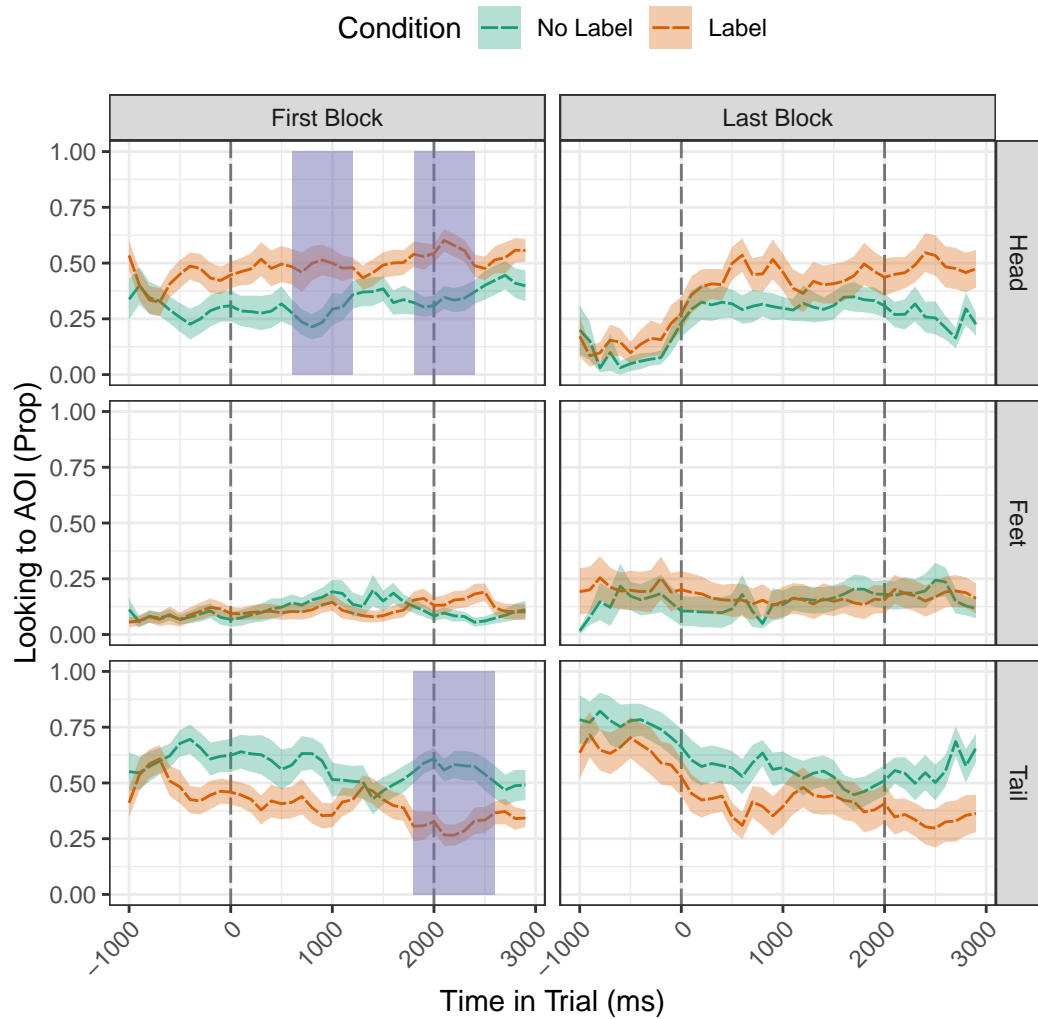


Figure 4: Time-course plot of the mean and SE of proportion looking at the different AOIs. Vertical dashed line represent feedback onset (0 ms) and labelling phrase onset (2000 ms). Purple overlay rectangles represent clusters where the difference between conditions reached significance.

attention to the head and decreased attention to the tail as a result.

First Look (Time) Another hypothesis we formulated was that participants would look more quickly at the diagnostic features when in the label condition compared to participants who did not hear a label. To test this, we submitted the log-transformed time to first look at the tail and feet from trial onset to a linear mixed-effects model. The model included main effects of and interaction between FstLst (first block, last block), AOI (head, feet, tail), Diagnostic feature (both, feet, tail), and Condition (no-label, label). The model also included random intercepts and slopes for FstLst, AOI, Diagnostic, and their interactions, by participant; and random intercepts by visual stimulus and by auditory stimulus. A summary of the parameter estimates and results of the ANOVA analysis on this model can be found in Table 5. A raincloud plot of the data can be seen in Fig. 5.

Two effects were significant here. First, there was a significant main effect of AOI, with participants in the no-label condition looking more slowly at the feet and more quickly at the tail compared to the head, when both feet and tail were diagnostic, during the first block of training. Second, there was a significant FstLst:AOI interaction, with participants in the no-label condition looking much more quickly at the feet and more quickly at the tail compared to the head, when both feet and tail were diagnostic, during the last block of training relative to the first block of training. No other effects were found to be significant, meaning that participants in the no-label condition did not look significantly quicker or slower to either AOI depending on which AOI was diagnostic, during the first block or last block, and that participants in the label condition did not significantly differ from participants in the no-label condition in any way. As such, labels did not impact the first AOI participants looked at, and instead, all participants started looking first at the diagnostic tail as early as the first block of training, but needed more training to also look earlier at the diagnostic feet.

4 Discussion

In this paper, we aimed to study the potential effects of labelling on attentional processes during categorisation. More precisely, we wanted to test whether an auditory label could direct attention towards diagnostic features when those features were of low salience.

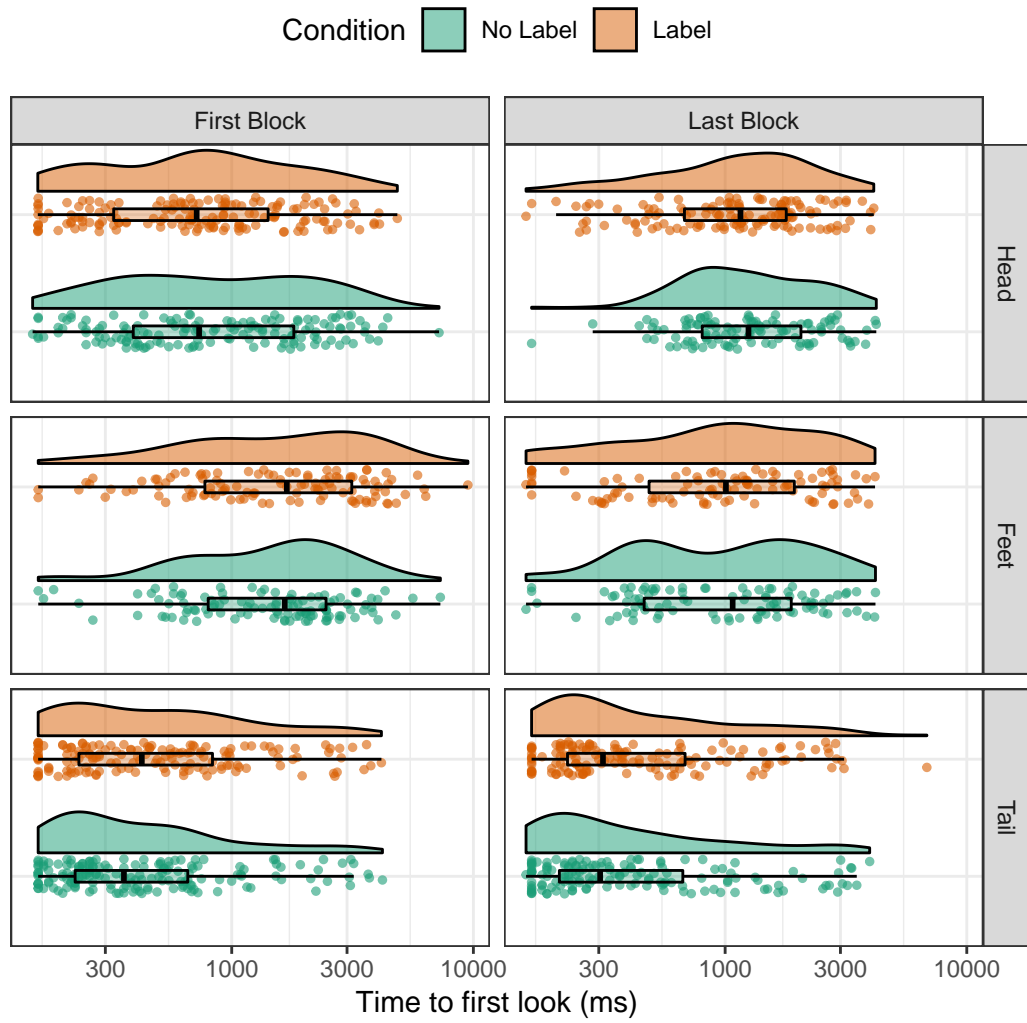


Figure 5: Raincloud plot from the data of the time before first look at each AOI.

| Parameter | Model Output | | ANOVA Output | |
|---|--------------|------------|--------------|-----------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 6.76 | 0.19 | | |
| FstLst | 0.49 | 0.22 | 1.18 | 0.287 |
| AOIFeet | 0.57 | 0.27 | 49.24 | .001 |
| AOITail | -0.56 | 0.27 | | |
| DiagnosticFeet | -0.03 | 0.16 | 0.12 | 0.888 |
| DiagnosticTail | 0.18 | 0.16 | | |
| Condition | -0.08 | 0.27 | 0.00 | 0.967 |
| FstLst:AOIFeet | -1.04 | 0.31 | 14.19 | .001 |
| FstLst:AOITail | -0.56 | 0.35 | | |
| FstLst:DiagnosticFeet | -0.13 | 0.24 | 0.32 | 0.727 |
| FstLst:DiagnosticTail | -0.09 | 0.23 | | |
| AOIFeet:DiagnosticFeet | 0.08 | 0.26 | 0.87 | 0.492 |
| AOITail:DiagnosticFeet | -0.16 | 0.25 | | |
| AOIFeet:DiagnosticTail | -0.30 | 0.28 | | |
| AOITail:DiagnosticTail | -0.43 | 0.23 | | |
| FstLst:Condition | -0.13 | 0.31 | 0.07 | 0.795 |
| AOIFeet:Condition | 0.00 | 0.38 | 1.03 | 0.369 |
| AOITail:Condition | -0.07 | 0.39 | | |
| DiagnosticFeet:Condition | -0.20 | 0.23 | 0.68 | 0.511 |
| DiagnosticTail:Condition | -0.23 | 0.22 | | |
| FstLst:AOIFeet:DiagnosticFeet | 0.32 | 0.38 | 0.55 | 0.700 |
| FstLst:AOITail:DiagnosticFeet | 0.30 | 0.37 | | |
| FstLst:AOIFeet:DiagnosticTail | 0.42 | 0.38 | | |
| FstLst:AOITail:DiagnosticTail | 0.29 | 0.36 | | |
| FstLst:AOIFeet:Condition | 0.58 | 0.43 | 0.11 | 0.900 |
| FstLst:AOITail:Condition | 0.32 | 0.49 | | |
| FstLst:DiagnosticFeet:Condition | 0.55 | 0.33 | 1.07 | 0.353 |
| FstLst:DiagnosticTail:Condition | 0.07 | 0.32 | | |
| AOIFeet:DiagnosticFeet:Condition | 0.34 | 0.37 | 1.50 | 0.219 |
| AOITail:DiagnosticFeet:Condition | 0.62 | 0.36 | | |
| AOIFeet:DiagnosticTail:Condition | 0.53 | 0.39 | | |
| AOITail:DiagnosticTail:Condition | 0.63 | 0.33 | | |
| FstLst:AOIFeet:DiagnosticFeet:Condition | -1.15 | 0.52 | 1.48 | 0.231 |
| FstLst:AOITail:DiagnosticFeet:Condition | -0.78 | 0.51 | | |
| FstLst:AOIFeet:DiagnosticTail:Condition | -0.72 | 0.53 | | |
| FstLst:AOITail:DiagnosticTail:Condition | -0.55 | 0.50 | | |

Table 5: Summary of the lmer model for first look time at the different AOIs during training.

First, we found no evidence of an effect of labelling on the behavioural level, with no significant differences between participants in the label and no-label condition in terms of learning speed or accuracy during training. In terms of looking patterns, however, we found that participants in the label condition preferred to look at the non-diagnostic head at the beginning of training. This head preference was no longer significant by the end of training. Thus, while the addition of an auditory label neither improved nor hindered categorisation itself, it did have an effect on attention distribution. Crucially, we did not replicate previous results in the adult literature on category learning, where the addition of an auditory label reduced reaction time and increased accuracy more quickly during training (e.g. Lupyan et al., 2007), and the effect we observed on looking patterns did not meet our predictions, specifically, participants did not look more at the diagnostic features (tail and feet) in the label condition.

First, we found that participants in both conditions looked much less at the feet than at any other feature, throughout training, and that they also took longer to make their first fixation at the feet than at any other feature. These results suggest that, although the tail and feet were equally diagnostic, the tail was naturally more salient than the feet, and even the feet's diagnosticity was not enough to make participants look at them as much as to the other diagnostic feature. Further, participants preferred to look at the tail compared to the head as early as the first block of training. We know however from previous research that animal heads are usually more salient than other features (Kovic et al., 2009), and we further know that for these particular stimuli the head was more salient than the tail for adults and infants (Capelier-Mourguy et al., 2019). Thus, participants here quickly learned that heads were not diagnostic, and consequently turned their attention to the next most salient feature: the tail. This points at the variety of studies that could be conducted, with a different number of features, with different salience relationships between them, and possibly differences in diagnosticity, to better understand how diagnosticity and salience interact in the presence or absence of category labels.

We further found an effect of labelling on attention. According to the *labels-as-symbols* theory, labels should highlight diagnostic features, thus helping to form categories. This has been confirmed in adults, in particular, auditory labels distort internal representation of categories to enhance the importance of diagnostic features, to the

detriment of other features (Lupyan, 2008). What we saw however was labels highlighting the already salient head, drawing attention away from the diagnostic tail and failing to increase attention to the other diagnostic feature, feet. This seemingly counterintuitive finding could result from the fact that, in the real world, categories are rarely defined by non-salient feature; in fact, it would make sense to believe that salient features have become salient because they were diagnostic for the categories we encounter in real life. Besides, studies have shown that labels reliably primed category representations (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012), which suggests that labels have *a priori* effects on attention and expectations. Thus, it would make sense that labels naturally activate representations in which features that would usually be diagnostic are highlighted, in this case the head, rather than highlight usually non-diagnostic features that are only diagnostic for the lab task at hand. This would also explain why the label only highlighted the head at the beginning of training, before participants learned that heads were not diagnostic for the current task. Moreover, in the first block of training, participants in the label condition first looked more at the head than participants in the no-label condition during the non-linguistic feedback. They then looked more at the head again, and less at the tail, later in the trial during the labelling phrase. This suggests that, knowing that the categories were labelled, participants expected the heads to be important for the categorisation feedback, both non-linguistic and linguistic. This is further evidence for global effects of auditory labels on participants' expectations about categories.

Additionally, this preference for heads in the label compared to the no-label condition could explain why we did not replicate another key finding in the literature. Auditory labels, even redundant, have been shown to improve categorisation performance (e.g. Lupyan et al., 2007). Here however, participants in both conditions needed the same number of training blocks to learn the categories, and their accuracy did not increase differently during training. Importantly, unlike in CMTW, we did not see here a clear ceiling effect, and thus this cannot alone explain the absence of difference. Yet, this result can be easily explained if we consider that auditory labels indeed helped participants form categories, but that this facilitatory effect was counteracted by the detrimental label-induced longer looking at the head at the beginning of training.

However, more work is needed to determine whether or not salient features in natural

categories are indeed diagnostic for those categories, and if labels always direct attention to those typically diagnostic features in lab tasks.

References

- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Rain-cloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. <https://doi.org/10.12688/wellcomeopenres.15191.1>
- Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PloS one*, *9*(7), e99670. <https://doi.org/10.1371/journal.pone.0099670>
- Althaus, N., & Plunkett, K. (2015). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, 1–11. <https://doi.org/10.1111/desc.12358>
- Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, *151*, 5–17. <https://doi.org/10.1016/j.jecp.2015.11.013>
- Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of experimental child psychology*, *64*(1), 3–26. <https://doi.org/10.1006/jecp.1996.2332>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of*

- Experimental Child Psychology*, 116(2), 105–119. <https://doi.org/10.1016/j.jecp.2013.05.002>
- Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2019). Learning categories with conflicting feature salience and diagnosticity. *PsyArXiv*.
- Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, 23(2), 178–186. <https://doi.org/10.1177/0956797611429133>
- Deng, W., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, 91, 24–62. <https://doi.org/10.1016/j.cogpsych.2016.09.002>
- Dink, J., & Ferguson, B. (2018). *eyetrackingR* [R package version 0.1.8]. R package version 0.1.8.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52. <https://doi.org/10.2307/2282330>
- Edmiston, P., & Lupyan, G. (2015). What makes words special? words as unmotivated cues. *Cognition*, 143, 93–100. <https://doi.org/10.1016/j.cognition.2015.06.008>
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2), 189–211. <https://doi.org/10.1080/19345747.2011.618213>
- Gelman, S. A., & Coley, J. D. (1991). Language and categorization: The acquisition of natural kind terms. In *Perspectives on language and thought: Interrelations in development* (pp. 146–196). Cambridge, Cambridge University Press.
- Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, 22(12), 2781–2789. <https://doi.org/10.1162/jocn.2010.21427>
- Graham, S. A., & Poulin-Dubois, D. (1999). Infants' reliance on shape to generalize novel labels to animate and inanimate objects. *Journal of child language*, 26(2), 295–320. <https://doi.org/10.1017/S0305000999003815>
- Kovic, V., Plunkett, K., & Westermann, G. (2009). Eye-tracking study of animate objects. *Psihologija*, 42(3), 307–327. <https://doi.org/10.2298/PSI0903307K>

- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.
- Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348–369. <https://doi.org/10.1037/0096-3445.137.2.348>
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. <https://doi.org/10.1111/j.1467-9280.2007.02028.x>
- Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception & Psychophysics*, *72*(8), 2236–2253. <https://doi.org/10.3758/APP.72.8.2236>
- Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186. <https://doi.org/10.1037/a0024904>
- Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06
- Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645. <https://doi.org/10.1037//0012-1649.36.5.635>
- Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language*, *135*, 66–72. <https://doi.org/10.1016/j.bandl.2014.05.005>
- Perry, L. K., & Lupyan, G. (2016). Recognising a zebra from its stripes and the stripes from "zebra": The role of verbal labels in selecting category relevant information. *Language, Cognition and Neuroscience*, 1–19. <https://doi.org/10.1080/23273798.2016.1154974>

- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681. <https://doi.org/10.1016/j.cognition.2007.04.003>
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Salverda, A. P., & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1122–1133. <https://doi.org/10.1037/a0023101>
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584. <https://doi.org/10.1146/annurev.ps.46.020195.003021>
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166–188. <https://doi.org/10.1037/0096-3445.133.2.166>
- Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, *111*(1), 65–86. <https://doi.org/10.1016/j.jecp.2011.07.007>
- Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child development*, 1695–1709. <https://doi.org/10.1111/1467-8624.00373>
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391–20120391. <https://doi.org/10.1098/rstb.2012.0391>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Chapter 5

A Model of Labelling and Attentional Focus

Using a new design, we were able in the previous chapter to avoid the ceiling effect observed in adults in Chapter 3. Doing so, we were able to detect differences in looking patterns induced by the addition of a redundant auditory label. According to previous studies on the effect of labels on categorisation in adults, we expected participants to look more at the non-salient but diagnostic features (tail and feet), but less at the salient but non-diagnostic head. Instead, we found evidence that participants, when hearing a label, looked reliably more at the head at the beginning of training, compared to participants who only heard non-linguistic feedback.

This finding, confusing at first, could be explained by considering participants' background knowledge. In the real world indeed, animal heads are arguably often diagnostic, and thus participants' looking behaviour at the beginning of training could reflect this background knowledge. It remains surprising, however, that participants in the label condition, at the end of training, still did not look more at diagnostic features compared to participants in the control group. Here again, it might be that looking patterns do not tell us the whole story about learning.

Overall, our empirical work in the last two chapters suggests that (a) 15-month-old infants see labels as object features, (b) adults have background knowledge linking auditory labels to animal heads in general, but most importantly (c) eye movement and looking time measures do not give us a good insight into learning mechanisms in these cases. To better understand how labelling might impact the learning of categories where the non-salient features, but not the salient features, are diagnostic, we propose to use computational modelling. In the next chapter, we describe a neurocomputational model simulating the attention bias induced by feature salience, and study how this model learns categories with and without a label.

Modelling the Interaction Between Auditory Labels and Attentional Focus

Arthur Capelier-Mourguy, Katherine E. Twomey, Gert Westermann
Department of Psychology, Lancaster University (UK)

Abstract

How labels relate to perceptual features of objects in category learning has been discussed controversially. According to one view labels have the same status as other features and become integrated into the object representation. Another view holds that labels are separate from object features and thereby shape object representations. Here we extended a previous computational model of object categorisation to model different ways in which labels can affect attention to object features during category learning. Specifically, we were interested in how object labels can direct learners' attention to diagnostic features of low salience, and replicated an empirical study recently designed to study this aspect of the question. Attention was modelled as modulation of learning rates of attended-to features. We discuss how changes in attention affect resulting object representation and the implications of these processes for the theories of the status of object labels in categorisation.

Keywords: connectionist model, representational development, label status, language development, cognitive development

1 Introduction

How labels relate to perceptual features of objects in category learning is controversial. It is clear from the literature that labels facilitate categorisation (e.g. Althaus & Westermann, 2016; Gelman & Coley, 1991; Gliga et al., 2010; Graham & Poulin-Dubois, 1999; Plunkett et al., 2008), but different mechanisms have been proposed to explain this effect, with no conclusive evidence in favour of a particular theory so far. On the *labels-as-symbols* account (Waxman & Gelman, 2009; Waxman & Markow, 1995), labels are symbolic, conceptual markers acting as privileged, top-down indicators of category membership, and label representations are qualitatively different to object representations. This implies that labels should shape the way we divide our attention when encountering an object, directly highlighting the relevant, diagnostic features for categorisation. Conversely, the *labels-as-features* account (Sloutsky & Fisher, 2004) considers labels to be equivalent to other (physical) features, and thus to be embedded into object representations. This theory does not predict any specific effects of labels on attention during categorisation. A third approach takes a middle ground between the *labels-as-symbols* and *labels-as-features* views: the *compound-representations* account (Westermann & Mareschal, 2014) acknowledges that language is a special kind of input and

that labels do not act at the same level as other features, but assumes integrated object representations are formed through the association between perceptual object features and labels. Although this theory does not make any explicit assumptions about attentional mechanisms, it assumes that labels will at first drive categorisation by adding to the overall similarity between exemplars of a category, and will become more closely related to object representations over time and make categories better defined by reducing the distance in representational space between exemplars of the same category. This, in turn, could optimise attention and/or reduce the cost of processing, when encountering new exemplars of known categories.

Numerous studies have addressed the question of the role of auditory labels on categorisation in infants, finding conflicting evidence. In support of the *labels-as-symbols* account, studies have shown that adding a label specifically allows infants to form categories that they would not otherwise form, for example grouping two different set of items into one category (Plunkett et al., 2008), or grouping a set of similar objects into two categories (Althaus & Westermann, 2016). More recent work using eye-tracking has shown that labels directed 8- to 12-month-old infants' attention to features of lower variability in a one-category categorisation task (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015). Conflicting with these findings, it has been shown for example that 4- to 5-year-old children and adults, when asked to make an inference on a missing feature of one of two previously learned categories, did not rely on the label if it was inconsistent with most of the other features (Deng & Sloutsky, 2012). More recently, 10-month-old infants displayed longer looking times when presented in silence with a previously labelled object than with a previously unlabelled object (Twomey & Westermann, 2017b), which was best explained as a novelty effect similar to what would be expected if a feature other than the label was missing from the object (Capelier-Mourguy et al., 2018).

In a recent empirical study, we extended previous work by training 15-month-old infants to categorise animals where the diagnosticity and salience of features mismatched (Capelier-Mourguy et al., 2019, thereafter CMTW). More precisely, we presented infants with two-featured snake-like animals with the head as a salient feature, and a less salient tail. The set of stimuli used can be seen in Fig. 1. Each feature was derived from one of two distinct exemplars, and pairs of features were then combined into novel, snake-like animals such that each type of head was paired equally often with both types of tail, and vice versa. For half of the infants, animals were given one of two names such that the type of tail varied consistently with the auditory label, but each type of head was heard equally often paired with each label; in other words, the low-salience tail was diagnostic for categorisation, and the high-salience head was not. As such, this study was the first to our knowledge to simultaneously control for the salience of different object features and their diagnosticity in a two-category categorisation task on infants. Infants were subsequently tested for successful recollection of the different features, and for word-learning in the relevant group. Testing the hypothesis that labels drive attention towards diagnostic features, we predicted that infants who heard labels would over time switch their attention away from the salient head to focus more on the diagnostic tail.

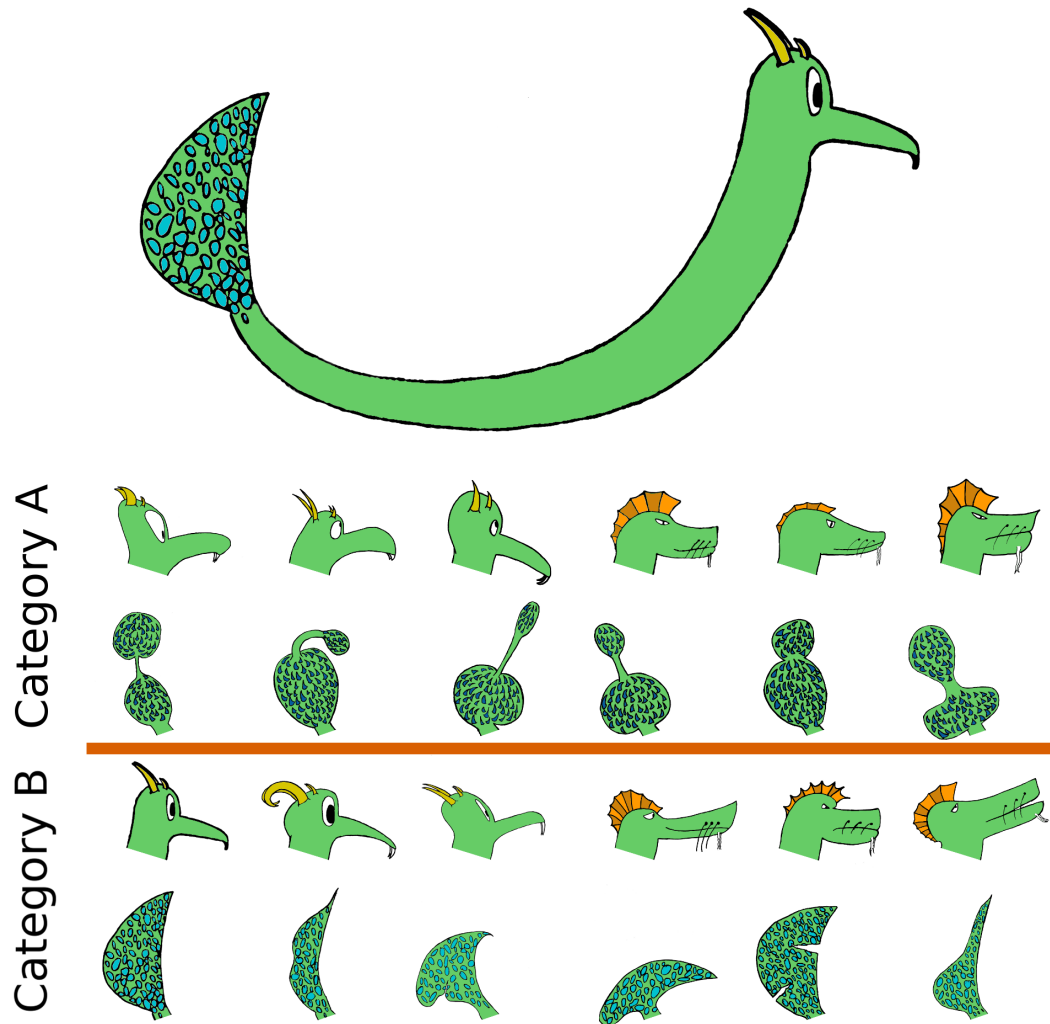


Figure 1: Example of a stimulus used for categorisation, and pairs of stimuli used during the familiarisation phase, in Capelier-Mourguy, Twomey, & Westermann (2019, thereafter CMTW).

However our predictions were not upheld: instead, infants' looking behaviour did not differ significantly depending on whether or not they heard a label, despite infants in the label group successfully learning the names and matching categories.

As such, these data conflict with the *labels-as-symbols* theory, whose main claim was that labels drive attention to diagnostic features. On the other hand, the *labels-as-features* view predicts no effects of labelling on attention, thus a true null in our empirical work would support this theory. However, the strong salience of the head in our experiment could have overshadowed any label-driven attention mechanisms, when previous studies supporting the *labels-as-symbols* theory used stimuli for which all features were uniformly salient. Here, we decided to use computational modelling to understand whether differences in feature salience between studies could explain the conflicting results observed. Specifically, we expect that labels will interact with features of different salience, and thus have a different impact on categorisation depending on feature

saliency.

Many models have been used to study categorisation and the effect of labels on categorisation (e.g. Althaus & Mareschal, 2013; Capelier-Mourguy et al., 2018; Erickson & Kruschke, 1998; Kruschke, 1992; Love et al., 2004; Mareschal & French, 2000; Mareschal et al., 2000; Mayor & Plunkett, 2010; Samuelson et al., 2011; Sloutsky & Fisher, 2004; Twomey & Westermann, 2017a; Westermann & Mareschal, 2014). Kruschke (1992) proposed the first computational model of categorisation that included an explicit attention mechanism: ALCOVE. This model combined an exemplar representation with perceptron-inspired error-driven back-propagation learning. Input neurons represented an explicit psychological dimension each, for example, a neuron could code for stimulus size, another one for brightness, and so on. Hidden neurons were previously encountered exemplars, with receptive fields in the multidimensional psychological representation space connecting them to the input neurons; as such, a new stimulus presented to the model would activate previously encountered exemplars depending on how similar they were to the new stimulus. Finally, output neurons represented the different categories to learn, making ALCOVE a supervised learning model. Crucially, each input node was gated by a dimensional attention strength, whose direct effect was to shape the receptive field of all hidden neurons over the corresponding psychological dimensions, allowing the model to learn on which dimensions to focus for a particular categorisation task. However, this implementation of an attention mechanism could not account for attention distribution over different features of a stimulus regardless of particular psychological dimensions. Furthermore, the use of explicit psychological dimensions restricts the model to experimenter choices on those dimensions.

Aside from models based on ALCOVE (Erickson & Kruschke, 1998; Love et al., 2004), no other model of human categorisation has implemented an explicit attention mechanism to our knowledge. This motivated us to develop a categorisation model with an explicit attention mechanism that would allow us to simulate infants' processing of stimuli with a known uneven saliency distribution, and test how this impacted their ability to learn new categories. To do so, we used a simple auto-encoder, and expanded it with a simple, theoretically plausible attention mechanism.

2 Methods

2.1 Model Architecture

Neurocomputational models have successfully captured looking time data from infant categorisation tasks (e.g. Mareschal & French, 2000; Twomey & Westermann, 2017a; Westermann & Mareschal, 2004, 2012, 2014). Here, we used a simple three-layer auto-encoder model to reproduce and explain data from CMTW. Auto-encoders reproduce input patterns on their output layer by comparing input and output activation after presentation of training stimuli, then using this error to adjust the weights between units using back-propagation (Rumelhart et al., 1986). In infant studies, looking times have been linked to information processing, with more complex or novel stimuli eliciting

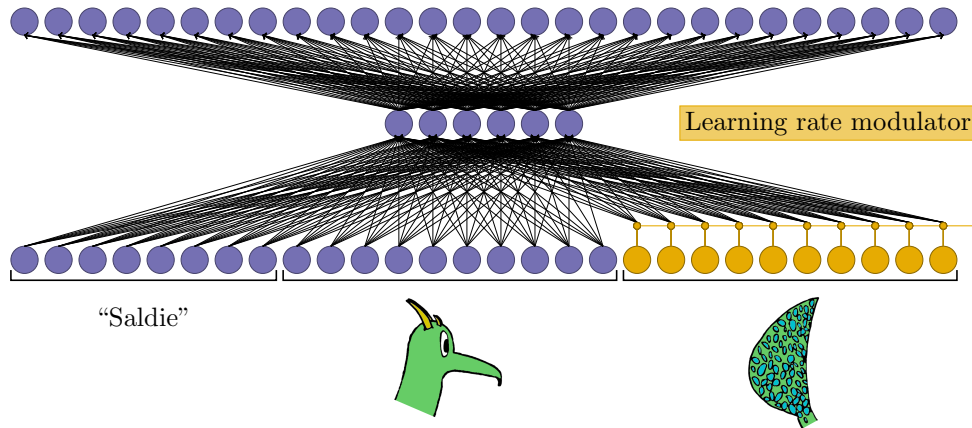


Figure 2: Structure of the attention-biased auto-encoder. The example stimuli at the bottom serve as an illustration but were not directly inputted to the model.

longer looking times (e.g. Houston-Price & Nakai, 2004; Oakes, 2010). As such, network error from auto-encoders has been used as a proxy for infant looking times.

A critical parameter for auto-encoders is their learning rate, which specifies how much the connection weights will be updated depending on the network error, or in other words, how much the model will learn from each presentation. Following the idea that a greater attention to salient features would lead to better encoding of those features in infants, we chose to implement salience in our model as a modulation in learning rate. More precisely, the learning rate of connections from the non-salient input to the hidden layer was reduced compared to the learning rate for all other connections. Thus, in the same way that infants will learn less from features they pay less attention to, our model will learn less from the non-salient feature than from any other feature at every step. We represent this as a ‘learning rate modulator’ on the network structure depicted in Fig. 2. The overall learning rate for the network was set at 0.01, and we ran models with the salience ratio between the low-salience tail and the rest of the network ranging from 10% (*i.e.* a learning ratio of $0.01 \times 10\% = 0.001$ for the tail input to hidden layer connections) to 90%, by increments of 10%.

2.2 Stimulus Encoding

Our stimuli were encoded as sets of abstract binary features that were designed to reflect the visual and label characteristics and the category structure of the stimuli used in CMTW. Thus, our encoding can be interpreted as a list of dummy variables that could generalize to alternative stimuli, coding for the presence/absence of one particular dimension of the stimuli (e.g. “has turquoise parts”, “has round shapes”, would be plausible dimensions for the stimuli considered here).

Each visual feature (head, tail) was encoded over ten units. Each visual feature existed in two versions, each built around a prototype by adding noise to it. More precisely, each feature exemplar was created by adding values drawn from a uniform distribution between -0.5 and 0.5 to the corresponding prototype, checking that there was a minimum distance between any two exemplars of the same category. The two

prototypes for each visual features had two overlapping units to represent the between-category similarities: for example, both heads and both tails in CMTW were partly green, and had a similar size.

The two labels were encoded over eight units with no overlapping units. In the no-label condition, all label units were simply set to 0.

Each stimulus was then built by combining a label with a head and a tail, following the structure used in CMTW and depicted in Fig.1. Notably, in the label condition, each type of tail was always associated with the same label, whereas there was an equal number of each type of head associated with each label, making the tail (and its associated label) fully and solely diagnostic for category membership and the head non-predictive.

2.3 Procedure

To collect an amount of data consistent with infant studies, we ran a total of 48 models for each of our ten salience ratios, 24 in each condition (no-label, label). We occasionally refer to independent models as ‘subject models’ in the rest of this paper.

In line with CMTW, the procedure consisted of a familiarisation phase, followed by contrast test trials in which the extent to which model subjects had encoded each feature was tested. Unlike in CMTW, we did not add word recognition trials at the end, as it was clear that the models in the label condition would have perfectly learned to match each label with the appropriate tail-based category.

Familiarisation The familiarisation phase lasted for 20,000 blocks. During each block, models were presented once with each one of the 12 stimuli in a pseudo-randomised order with exemplars from each category alternating. The first category presented for each model was randomised. Network error and hidden representations for each stimulus were recorded every 50 blocks.

Contrast Test Trials In CMTW, contrast test trials consisted of two animals presented side by side in silence, one with new versions of a familiarised head and tail (thereafter “old features”), and one with an old feature and a new type of head or tail (for head and tail contrast test trials). The prediction in developmental psychology is that, if infants have fully encoded a feature, for example the tail, then they will exhibit preferential looking towards the new tail compared to the old tail; that is, infants will show a *novelty preference*.

To reproduce this procedure in our models, we considered the number of successive presentations of the same stimulus necessary for the network error to fall below a predefined threshold of 10^{-2} , or for a maximum of 200 iterations. To compare this measure between an old stimulus and a stimulus with a new feature, we saved the model state after familiarisation and presented this saved state with each stimulus. Crucially, stimuli in the contrast test trials were always presented without a label. The assumption was that if labels during familiarisation enhanced learning of the tail units, subsequent

presentation of a stimulus with a similar tail would be encoded faster than a stimulus with a new kind of tail. Conversely, a label would not be expected to enhance learning of the head units (as head types did not systematically co-occur with specific labels), thus presenting at test two stimuli with a familiar head against a new head might not lead to different speeds of encoding.

2.4 Data Handling and Software Specifications

Data Handling A common measure used with auto-encoders is the network error on the output layer, which has been used as a proxy for looking time (e.g. Mareschal & French, 2000; Twomey & Westermann, 2017a; Westermann & Mareschal, 2012, 2014). Here, we recorded the network error separately over label, head, and tail units, as a proxy for looking time to the two visual AOIs and processing of the auditory information.

Another measure of interest with neural networks in general is their hidden layer, providing an insight into the model’s internal representations of items it has encoded (e.g. Mareschal & French, 2000; Rogers & McClelland, 2004; Westermann & Mareschal, 2012, 2014). This measure allows us to better understand how the model learns to group objects into categories.

Here, we first used Principal Component Analysis (PCA) to reduce the dimensionality of the representational space of the hidden layer (activation pattern over all hidden units in response to a specific input) in order to plot the 6-dimensional representations in a 2-dimensional space. We ran an independent PCA for each subject for each block (first and last). The direction of each axis in a PCA being random, we then changed the sign of each PCA so that the average tail A would always be in the top-right quadrant of the plot; this ensured that any hypothetical clusters would be consistently positioned across participants.

We then computed the average absolute within-category distance for each category (mean pairwise distance between all exemplars in a category), the between-category distance as the distance between the cluster centre for each category, and the average relative within-category distance as the absolute within-category distance divided by the between-category distance. This is important as, over time, the model learns to differentiate between exemplars of a category, increasing the absolute within-category distance, but also learns to bring each category into distinct clusters. That is, the between-category distance increases more than the within-category distance, making the categories relatively more compact.

Finally, for contrast test trials, we recorded the total number of presentations of each stimulus necessary for the network error to fall below a predefined threshold, or for a maximum number of 200 presentations. We then computed the novelty preference for head and tail contrast trials by dividing the number of presentations for the “new” stimulus by the summed number of presentations for the new and old stimuli. To account for the boundedness of this proportion of looking at the new stimulus, we then ran statistics on the arcsine-root transformation of this measure. Crucially, we did not here use network error as a proxy for looking time. This is because, due to the different

learning rates for the salient head and non-salient tail, different network errors might here lead to slower or quicker encoding depending on how the error is distributed over the different units.

Software Specifications All source code and data are available online¹. Simulations were run using Python (version 3.6.8) and `numpy` (version 1.13.3). All statistical results were obtained using R (version 3.5.2). Analyses in this paper were conducted using (a) `lme4` (version 1.1-17; Bates et al., 2015) to run Sample Theory Based (STB) mixed-effects models and `lmerTest` (version 3.0-1; Kuznetsova et al., 2017) to run ANOVA analyses on those mixed-effect models, and (b) `ggplot2` (version 2.2.1; Wickham, 2016) to plot graphs from our data and `ggeffects` (version 2.4.1; Lüdtke, 2018) to compute and plot estimated marginal effects from our models.

3 Results

Analysis Structure For our statistical analyses, we used (generalised) linear mixed-effects models as implemented in `lme4` fitted with maximal converging random-effects structure to estimate parameters (Barr et al., 2013). For significance testing of those parameters, we used type I ANOVA analyses with Satterthwaite’s method as implemented in `lmerTest` for linear models, and commonly-used asymptotic Wald tests for generalised linear models.

3.1 Familiarisation

3.1.1 Looking Times

We submitted network error (looking time) to both visual features (head and tail) to a linear mixed-effects model. The model included main effects of and interaction between scaled block number (`z.block`), condition (no-label, label), `error_type` (salient feature, non-salient feature), and `saliency_ratio`. The model also included random intercepts and slope for scaled block, `error_type`, and their interaction, by subject model. A summary of the model’s parameter estimates and ANOVA results for those parameters are given in Table 1. A time course plot of the data for small saliency ratio (tail 20% as salient as the head), medium saliency ratio (50% as salient), and high saliency ratio (80% as salient), for each feature, is shown in Fig. 3.

Looking both at the parameter estimates and the plot, the most notable results were the main effect of `error_type` and the condition-by-`error_type`, and their interaction with `z.block`, with a much higher error for the non-salient tail in the no-label condition throughout learning for low saliency ratios. Other interesting results were the main effect of condition and the `z.block`-by-condition interaction, with smaller error for the salient head in the label condition, or in other words, less learning in the no-label condition for the head, despite its high saliency.

¹<https://github.com/respatte/SaliencyDiagnosticityEmpirical>

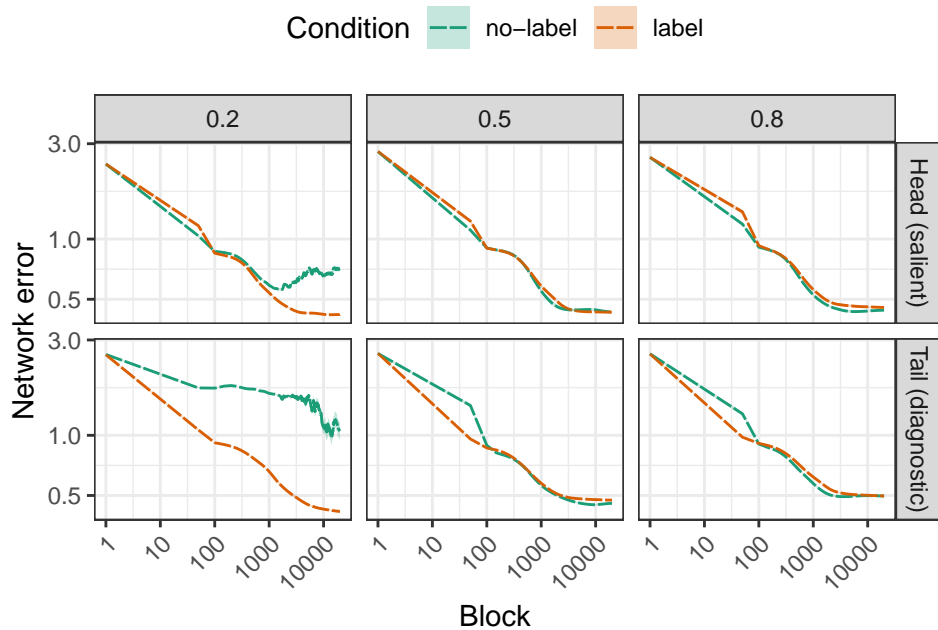


Figure 3: Time-course plot of the mean and SE of network error (looking time) on each feature, for different salience ratios.

| Parameter | Model Output | | ANOVA Output | |
|---|--------------|------------|--------------|------------------------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 0.63 | 0.007 | | |
| z.block | -0.05 | 0.006 | 705.97 | $< 2.2 \cdot 10^{-16}$ |
| condition | -0.13 | 0.009 | 77.31 | $< 2.2 \cdot 10^{-16}$ |
| error_type | 0.78 | 0.028 | 170.11 | $< 2.2 \cdot 10^{-16}$ |
| salience_ratio | -0.13 | 0.011 | 149.36 | $< 2.2 \cdot 10^{-16}$ |
| z.block:condition | 0.02 | 0.008 | 7.53 | .00631 |
| z.block:error_type | -0.29 | 0.017 | 97.07 | $< 2.2 \cdot 10^{-16}$ |
| condition:error_type | -0.63 | 0.040 | 49.66 | $7.35 \cdot 10^{-12}$ |
| z.block:salience_ratio | 0.01 | 0.010 | 65.36 | $6.47 \cdot 10^{-15}$ |
| condition:salience_ratio | 0.22 | 0.016 | 186.27 | $< 2.2 \cdot 10^{-16}$ |
| error_type:salience_ratio | -1.09 | 0.050 | 218.39 | $< 2.2 \cdot 10^{-16}$ |
| z.block:condition:error_type | 0.24 | 0.024 | 41.70 | $2.88 \cdot 10^{-10}$ |
| z.block:condition:salience_ratio | -0.01 | 0.014 | 33.23 | $1.57 \cdot 10^{-8}$ |
| z.block:error_type:salience_ratio | 0.39 | 0.030 | 115.07 | $< 2.2 \cdot 10^{-16}$ |
| condition:error_type:salience_ratio | 0.92 | 0.071 | 114.72 | $< 2.2 \cdot 10^{-16}$ |
| z.block:condition:error_type:salience_ratio | -0.33 | 0.043 | 60.04 | $6.81 \cdot 10^{-14}$ |

Table 1: Parameter estimates and ANOVA results for the STB model on network error (looking time) on both visual features.

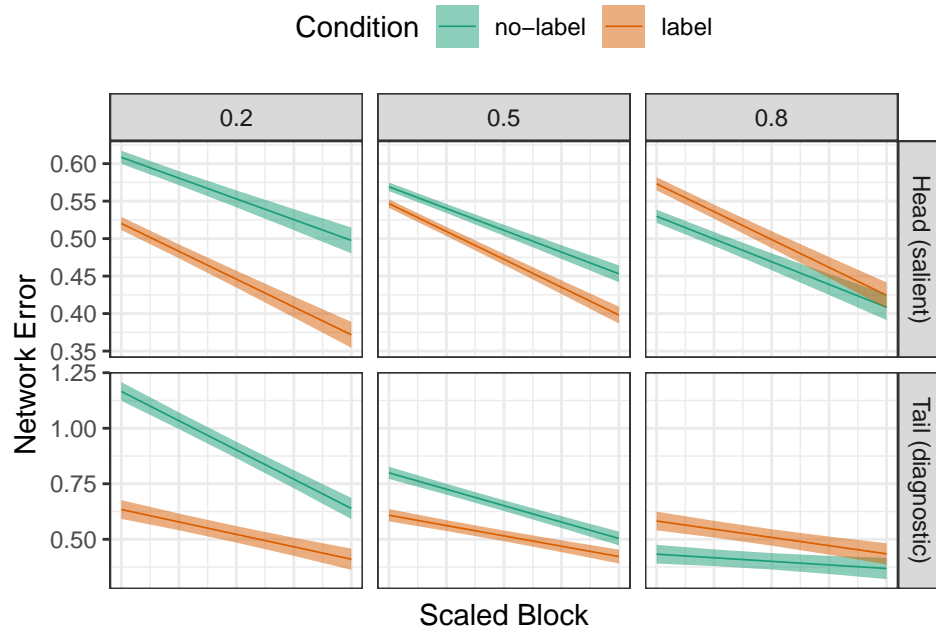


Figure 4: Estimated marginal effect of `z.block` for network error during learning on each feature and condition, for different salience ratios.

All other parameters of the model were also significant, but their main impact in this model was to cancel out those differences for low salience ratios, since network errors between conditions and features were similar for higher salience ratios, as can be seen in the marginal effects plot in Fig. 4.

3.1.2 Hidden Representations

We submitted the average relative within-category distance to a linear mixed-effects model. The model included main effects of and interaction between scaled block number (`z.block`), condition (`no-label`, `label`), and `salience_ratio`. The model also included random intercepts and slope for scaled block by subject model. A summary of the model's parameter estimates and ANOVA results for those parameters are given in Table 2. A plot of the first two dimensions of a PCA on the hidden representations for the first and last block of learning is shown in Fig. 5, and a time course plot of the data for small salience ratio (tail 20% as salient as the head), medium salience ratio (50% as salient), and high salience ratio (80% as salient) is shown in Fig. 6.

The main effects of condition and `salience_ratio` were significant, with subject models in the `no-label` condition having a higher relative within-category distance for low salience ratios, but this within-category distance decreasing as salience ratio increased.

The condition-by-`salience_ratio` interaction was also significant, its main effect being to keep the relative within-category distance comparable across different salience ratios in the `label` condition, with no difference between the two conditions for high salience ratios, as can be seen in the marginal effects plot in Fig. 7. Thus, category compactness was specifically impaired in the `no-label` condition for small salience ratios, but increasing

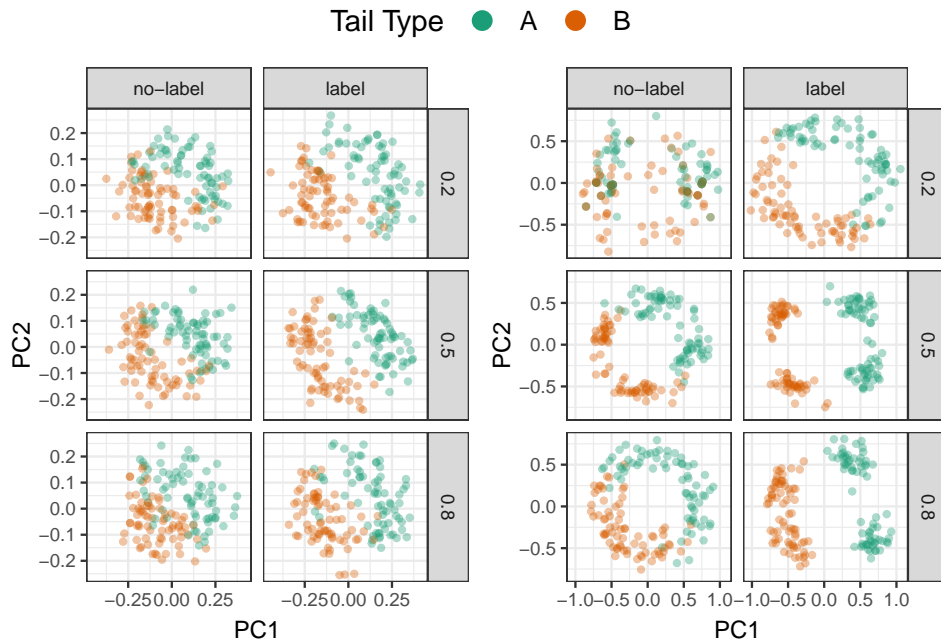


Figure 5: Hidden representations of items grouped by tail type for the first block of learning (left) and last block of learning (right) (first two dimensions of a PCA).

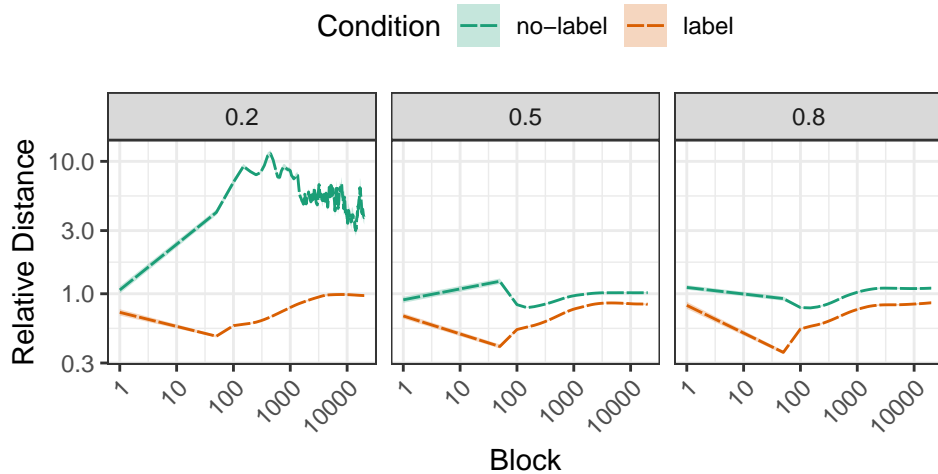


Figure 6: Time-course plot of the mean and SE of average relative within-category distance, for different salience ratios.

| Parameter | Model Output | | ANOVA Output | |
|----------------------------------|--------------|------------|--------------|------------------------|
| | Estimate | Std. Error | F value | $Pr(> F)$ |
| (Intercept) | 4.20 | 0.170 | | |
| z.block | -0.05 | 0.068 | 18.69 | .613 |
| condition | -3.27 | 0.240 | 80.84 | $< 2.2 \cdot 10^{-16}$ |
| saliency_ratio | -4.40 | 0.302 | 71.98 | $< 2.2 \cdot 10^{-16}$ |
| z.block:condition | 0.08 | 0.096 | 40.67 | .457 |
| z.block:saliency_ratio | 0.08 | 0.120 | 56.43 | .725 |
| condition:saliency_ratio | 4.29 | 0.427 | 71.83 | $< 2.2 \cdot 10^{-16}$ |
| z.block:condition:saliency_ratio | -0.09 | 0.170 | 56.70 | .578 |

Table 2: Parameter estimates and ANOVA results for the STB model on average relative between-category distance.

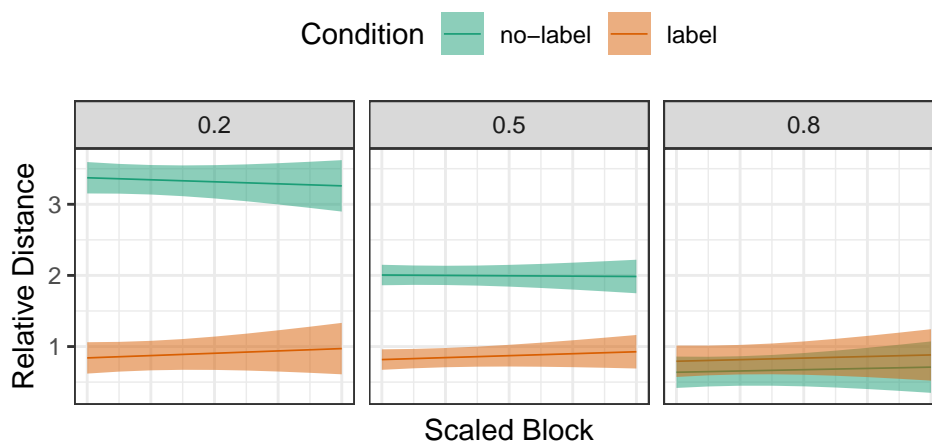


Figure 7: Estimated marginal effect of z.block for relative between-category distance on each feature and condition, for different saliency ratios.

saliency ratios reduced this impairment to the point that category compactness did not differ between conditions.

3.2 Contrast Test Trials

We submitted chance-corrected novelty preference to a linear mixed-effects model. The model included main effects of and interactions between condition (no-label, label), contrast_type (head contrast, tail contrast), and saliency_ratio. The model also included random intercepts by subject. A “raincloud” plot (Allen et al., 2019) of the data is shown in Fig. 8. These plots include a half-violin plot to understand the shape of the data, individual data points to better understand the structure of the data, and a boxplot to give some descriptive statistics at-a-glance.

Since we were interested in knowing whether or not there was a novelty preference in either condition in either contrast trial for all saliency ratios, we computed chance-corrected estimated marginal means and 95% confidence intervals for those variables and display those values in Fig. 9.

It is clear from this plot that there were no differences between conditions. The

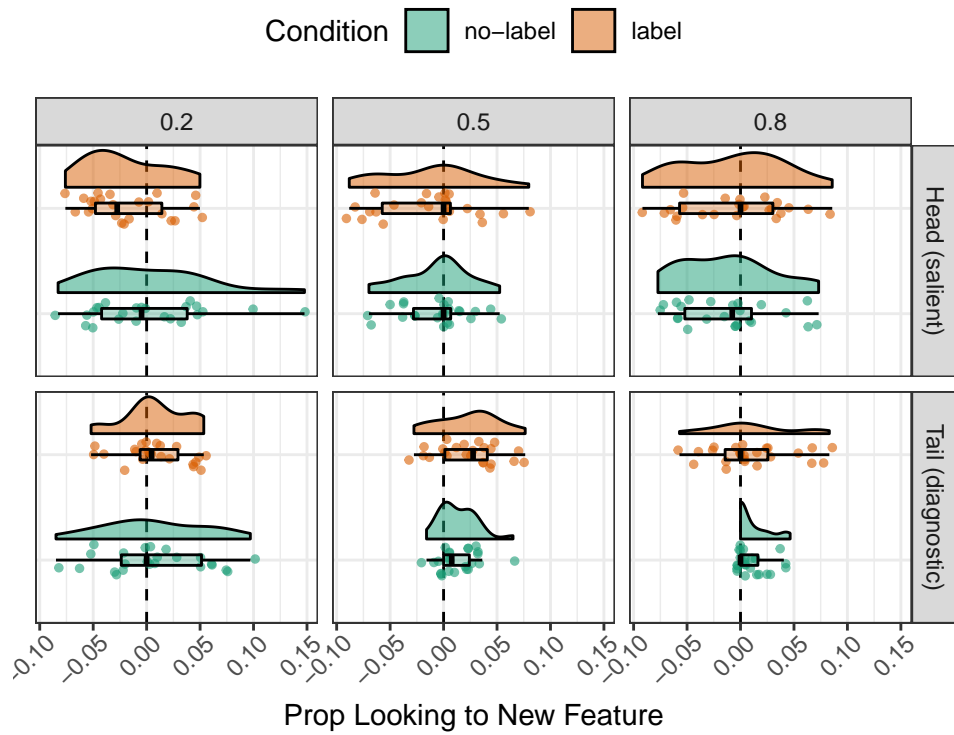


Figure 8: Raincloud plot of the “proportion of looking” at the stimulus with a new feature compared to the stimulus with only old features for head and tail contrast trial.

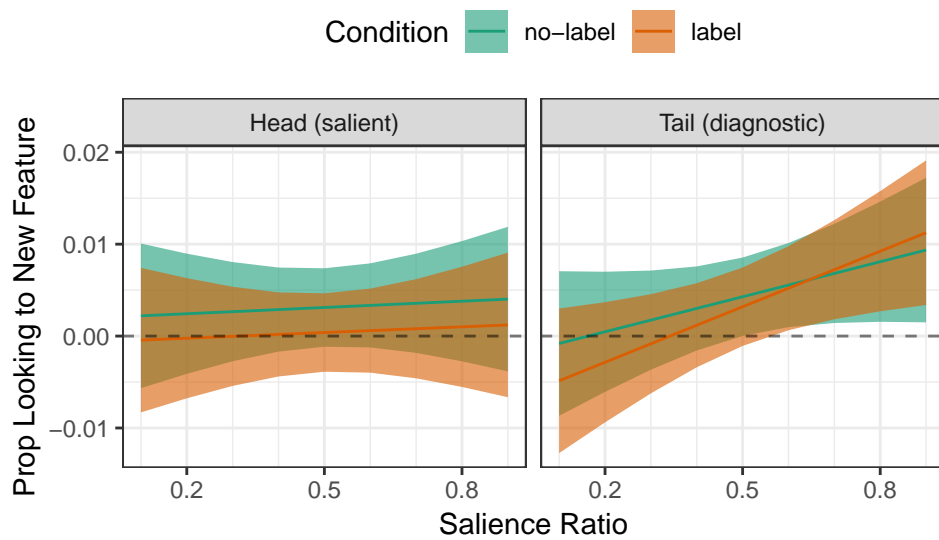


Figure 9: Estimated marginal effect of salience ratio by condition and contrast type. Ribbons represent 95% confidence intervals.

only other clear result is that the models required a comparable number of iterations to encode both stimuli in the head trials, and only needed slightly more simulations to encode the stimulus with a new feature in the tail contrast trials for higher salience ratios, with no difference depending on label condition. However, this difference was very small in magnitude and thus does not really warrant further interpretation, as such a small effect does not compare to effects observed in infant data, and would not be detectable with empirical work on infants.

4 Discussion

In this paper, we proposed a new implementation of feature salience for a categorisation auto-encoder model, and use it to test the interaction between labelling, feature diagnosticity, and salience, in a categorisation task, following an empirical design asking the same question. Specifically, we considered feature salience as an attentional bias towards specific item features, and implemented the impact of this attentional bias as a difference in learning rates, with smaller learning rates for features of lower salience. We then studied the impact of labelling when learning categories where a non-salient feature (here a tail) is solely diagnostic of category membership, depending on the difference in salience between this diagnostic feature and a more salient, non-diagnostic feature (here a head). Overall, we found that adding a label positively impacted learning during training, but that it did not have a strong effect on encoding of new within- and out-of-category exemplars in a subsequent test phase.

First of all, for low salience ratios, the model could not reduce its prediction error (an index of learning) to the low-salience tail in the no-label condition. Put differently, without a label, the low-salience feature was not well encoded. Conversely, the presence of a label allowed the model to encode the low-salience feature better. Crucially, although this result first seems compatible with a *labels-as-symbols* theory in which labels actively highlight diagnostic features, the label in our model was not different from other physical features; rather, the mere statistical co-variation between the label and the tail improved learning of the diagnostic tail. Thus, similar empirical results can be explained without the need to evoke an explicit label-induced attention driving mechanism. For example, the longer looking to and more robust encoding of a diagnostic feature in the presence of a label evidenced by Althaus and colleagues (Althaus & Mareschal, 2014; Althaus & Plunkett, 2015), does not necessarily entail an effect of the auditory label highlighting the diagnostic feature, but instead can be more simply explained in terms of statistical co-variation between the auditory label and diagnostic feature.

Second, for low salience ratios again, learning of the head without a label was impaired, even though the head was salient and not linked to the label in any way. With a label, as for the non-salient tail, the model learned the head rapidly. This can be explained by the fact that auto-encoders, much like humans, build internal representations for entire objects; thus, a difficulty in learning the tail induced by its low salience is reflected more generally in the inability to form a good representation of the encountered exemplars, and thus to learn efficiently the salient head.

These results were however not only dependent on which feature was salient and which was not, but on the salience difference between the two features. Specifically, for small to medium differences in salience, both features were equally well encoded in the presence or absence of a label. With a great salience discrepancy, however, without a label, our data suggests that there is a competitive process between the two feature to encode inputs (see Fig. 3, leftmost panels): as the error over the tail starts decreasing around the thousandth step, the error over the head starts increasing again. Overall, the salient head wins this competition, and since it is not informative for the encoding of the tail, this leads to a poorer encoding of the stimuli overall. The addition of a label, equally salient compared to the head, and informative for the tail, levels out the competition between the two features and thus allows the model to better encode the stimuli overall.

With respect to the contrast test trials, we found no strong evidence for a novelty or familiarity preference, with models in both conditions needing equally long to process a new exemplar from the familiarised categories or an exemplar displaying a novel, out-of-category feature, regardless of which feature (head or tail) was novel, and regardless of the salience ratio between the head and tail. This does not replicate looking time results from CMTW, in which infants in the no-label condition exhibited a strong novelty preference in the tail contrast trials, and infants in the label condition exhibited a mix of familiarity and novelty preferences in those same tail contrast trials. This might reflect the fact that our measure (number of iterations before network error fell under a predetermined threshold) is not a good proxy for infant looking times in this task, or be an indication that our choice of implementation for salience was incorrect. Importantly, we did not for this task use the typical network error. Instead, considering that the differences in learning rates might impact how our model learned from its error, we looked at the number of presentation needed to fully encoded the stimuli, a measure we believed to be a better proxy for learning. Further work is needed to assess if this measure can indeed provide a good understanding of learning and be a good proxy for infant looking times.

Further, our familiarisation results in terms of network error, typically used as a proxy for looking times, do not replicate empirical data from CMTW either. Specifically, feature salience in CMTW's stimuli differed greatly, yet they found no differences in looking patterns during familiarisation depending on label condition. On the opposite, we found that, for great salience differences, a label would have a positive impact on network error. Although this can be seen as evidence against our implementation of feature salience, this, taken together with previous work finding evidence of successful learning at test in infants without showing as systematic patterns of looking during training (e.g. Aslin, 2007; Hilton et al., 2019; Hilton & Westermann, 2017; Twomey et al., 2018), brings further evidence that looking times do not directly measure information processing. Rather, looking times are a proxy for those attention processes that impact object exploration, but do not measure the cognitive resources dedicated to each look. Thus, since network error in neurocomputational models is a clear measure of learning and not

only of attention and spatial exploration, then future work is needed to understand how exactly modelling results can relate to eye-tracking results, and in general more work is needed to understand what internal processes are showcased by looking patterns.

However, all those discrepancies between our modelling results and the empirical results in CMTW might be more easily explained if our implementation of feature salience was incorrect. Although our model does not aim to replicate directly brain processes but is merely an abstract representation of those processes, studies on the neurobiological bases of salience, that is, selective attention, can help us better understand what would be better candidates to model salience. The model that is generally accepted is that of a gain model, in which the firing rates of cells that respond to attended stimulus increases while the firing rates of cells responding to other unattended stimuli decrease (see Caporello Bluvas & Gentner, 2013, for a review). We first attempted a naive implementation of this system by directly using larger input values for salience features (and smaller input values for non-salient features), without success. Indeed, in our abstract model, the units do not represent biological neurons or neuron populations, and thus the activation values do not relate directly to firing rates. Arguably, our current implementation attempted to represent this same effect, with more plasticity and therefore more learning the enhanced salience features, but here again, this might not relate directly to enhancement and inhibition effects in the brain.

Crucially, these enhancing and inhibiting effects are thought to be a top-down process driven by a sustained representation of a model of behavioural event, possibly in the prefrontal cortex, which feeds back into lower level processes to enhanced neural activities relevant for the current event (Merzenich et al., 2014). One way to follow this neurobiological result to implement salience into our model would be to change how much discrepancies between the network output and its input will impact the backpropagated weight updates. In other words, instead of acting on the learning rate at the level on the input units, in a somewhat *bottom-up* way, we could attempt to change the learning rates between the output and hidden units, in a more *top-down* way.

In conclusion, even though it is unclear whether our implementation of feature salience for neurocomputational models is valid, our work, taken together with conflicting results in the literature, suggests that differences in salience between stimulus features in a categorisation task can impact how an auditory label may interact with the categorisation process. This might explain how empirical studies that differ with respect to how they build their stimuli and control for feature salience fail to reach a consensus on the effect of labels on categorisation in infants.

References

- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63. <https://doi.org/10.12688/wellcomeopenres.15191.1>

- Althaus, N., & Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Transactions on Autonomous Mental Development*, 5(4), 288–297. <https://doi.org/10.1109/TAMD.2013.2264858>
- Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PloS one*, 9(7), e99670. <https://doi.org/10.1371/journal.pone.0099670>
- Althaus, N., & Plunkett, K. (2015). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, 1–11. <https://doi.org/10.1111/desc.12358>
- Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, 151, 5–17. <https://doi.org/10.1016/j.jecp.2015.11.013>
- Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1), 48–53. <https://doi.org/10.1111/j.1467-7687.2007.00563.x>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2018). Neurocomputational models capture the effect of learned labels on infants' object and category representations. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1. <https://doi.org/10.1109/TCDS.2018.2882920>
- Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2019). Learning categories with conflicting feature salience and diagnosticity. *PsyArXiv*.
- Caporello Bluvás, E., & Gentner, T. Q. (2013). Attention to natural auditory signals. *Hearing Research*, 305, 10–18. <https://doi.org/10.1016/j.heares.2013.08.007>
- Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, 23(2), 178–186. <https://doi.org/10.1177/0956797611429133>
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127(2), 107. <https://doi.org/10.1037/0096-3445.127.2.107>
- Gelman, S. A., & Coley, J. D. (1991). Language and categorization: The acquisition of natural kind terms. In *Perspectives on language and thought: Interrelations in development* (pp. 146–196). Cambridge, Cambridge University Press.
- Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, 22(12), 2781–2789. <https://doi.org/10.1162/jocn.2010.21427>

- Graham, S. A., & Poulin-Dubois, D. (1999). Infants' reliance on shape to generalize novel labels to animate and inanimate objects. *Journal of child language*, *26*(2), 295–320. <https://doi.org/10.1017/S0305000999003815>
- Hilton, M., Twomey, K. E., & Westermann, G. (2019). Taking their eye off the ball: How shyness affects children's attention during word learning. *Journal of Experimental Child Psychology*, *183*, 134–145. <https://doi.org/10.1016/j.jecp.2019.01.023>
- Hilton, M., & Westermann, G. (2017). The effect of shyness on children's formation and retention of novel word–object mappings. *Journal of Child Language*, *44*(6), 1394–1412. <https://doi.org/10.1017/S030500091600057X>
- Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. <https://doi.org/10.1002/icd.364>
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. <https://doi.org/10.1037/0033-295X.99.1.22>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. <https://doi.org/10.1037/0033-295X.111.2.309>
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.
- Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06
- Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645. <https://doi.org/10.1037//0012-1649.36.5.635>
- Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31. <https://doi.org/10.1037/a0018130>
- Merzenich, M. M., Van Vleet, T. M., & Nahum, M. (2014). Brain plasticity-based therapeutics [Publisher: Frontiers]. *Frontiers in Human Neuroscience*, *8*. <https://doi.org/10.3389/fnhum.2014.00385>
- Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development*, *11*(3), 255–268. <https://doi.org/10.1080/15248371003699977>
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681. <https://doi.org/10.1016/j.cognition.2007.04.003>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space (J. Wiles, Ed.). *PLoS ONE*, *6*(12), e28095. <https://doi.org/10.1371/journal.pone.0028095>
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166–188. <https://doi.org/10.1037/0096-3445.133.2.166>
- Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*, 413–438. <https://doi.org/10.1111/cogs.12539>
- Twomey, K. E., & Westermann, G. (2017a). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, e12629. <https://doi.org/10.1111/desc.12629>
- Twomey, K. E., & Westermann, G. (2017b). Learned labels shape pre-speech infants' object representations. *Infancy*, *23*, 61–73. <https://doi.org/10.1111/inf.12201>
- Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. <https://doi.org/10.1016/j.tics.2009.03.006>
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Westermann, G., & Mareschal, D. (2004). From parts to wholes: Mechanisms of development in infant visual object processing. *Infancy*, *5*(2), 131–151. https://doi.org/10.1207/s15327078in0502_2
- Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, *27*(4), 367–382. <https://doi.org/10.1016/j.cogdev.2012.08.004>
- Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391–20120391. <https://doi.org/10.1098/rstb.2012.0391>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Chapter 6

General Discussion

In this thesis, we aimed to study the role of auditory labels in categorisation, through the use of empirical and computational modelling methods. Overall, we provided mild evidence in favour of the *labels-as-features* theory, stating that early in development infants treat auditory labels as object features. Importantly, labels do not have a preferential role according to this theory, rather, they help categorisation by increasing the similarity of different exemplars that all share the same label within a category. Crucially, we failed to find any evidence in favour of the *labels-as-symbols* theory. According to this theory, labels have from an early developmental stage a symbolic value as category markers, and help categorisation by highlighting commonalities between category exemplars, that is, diagnostic features. Further, we did not directly replicate results in the adult literature, neither in terms of facilitatory effects of redundant labels for category learning nor in terms of labels driving attention to diagnostic features. Instead, we found no differences in terms of learning speed, and labels drove attention to a salient but non-diagnostic feature. Finally, we evidenced category learning in the presence of a label in 15-month-old infants, with no differences in terms of looking patterns compared to infants in a control group.

First, we reproduced and extended a neurocomputational model of category learning to explain existing empirical data and help tease apart two theories that could both explain the observed data as resulting from different mechanisms. Specifically, in the empirical study, 10-month-old infants were familiarised at home with two objects, and only heard a label for one of them. In a subsequent lab task, in which infants were presented with pictures of the two objects one at a time, infants looked longer when seeing the previously labelled object. This could be explained by the *labels-as-features* theory as a novelty effect, due to the absence of one feature of the object, the label. Alternatively, the *compound-representations* theory expected that seeing the object in silence would activate its corresponding label, which would in turn increase infants' attention to the stimulus in the same fashion the presence of a label would. Implementing those theories into two structurally similar models allowed us to precisely test their predictions on the experimental design considered. Only our model implementing the *labels-as-features* theory reproduced the statistical effects observed in the data. In conclusion, for the experimental design we tested, our model suggested that 10-month-old infants treated

labels as object features, not as separate features nonetheless integrated into object representations. We further used this *labels-as-features* model to predict results for an ongoing follow-up empirical study by the same authors, in which infants were habituated with simple categories made of a few exemplars, rather than single objects. The model once predicted that infants, in a subsequent lab test in silence, would look longer at the previously labelled category. If the empirical results matched our prediction, this would corroborate our finding that 10-month-old infants view labels as object features.

We then tested a key prediction of the *labels-as-symbols* theory on 15-month-old infants, with adult participants as a control group that we knew sees labels as symbolic markers. This theory argues that labels can help infants group objects into categories by highlighting the defining, diagnostic features for those categories. To test this prediction, we used two-featured animal-like stimuli for which we knew one feature was more salient (the head), and deliberately grouped them into two named categories according to their less salient feature (the tail). We thus expected infants in a control group to preferentially look at the salient head, and infants who heard auditory labels to look more at the diagnostic tail in order to learn the correct categories. We expected similar results for adults, with the distinction that adults in both the label and no-label group were provided with non-linguistic feedback. Thus, following previous literature, we only expected adults who heard a redundant label to be quicker at learning the category and to look quicker and/or more at the diagnostic tail compared to adults who only heard non-linguistic feedback. Interestingly, none of our predictions were upheld. Specifically, labels failed to attract infants' attention to the diagnostic tail, but we nonetheless gathered evidence that infants did learn the correct label-category matching. Thus, the absence of an attention-driving mechanism did not lead to a failure to learn the categories. As such, our data provide indirect evidence against the *labels-as-symbols* theory, and consequently they provide indirect evidence in favour of the *labels-as-features* theory. Most importantly, the successful categorisation in the absence of any differences in terms of looking behaviour between infants who heard a label and those who did not, taken together with previous literature, suggests that eye-tracking data are not a proxy of learning in general, but specifically tell us about attention-driven exploration processes. What mechanisms impact these processes, and how different levels of attention when looking at different features of the world impact learning mechanisms, remains to be studied further.

In this study, we also failed to replicate previous findings in the adult literature (improved category learning and increased attention to diagnostic features, when hearing a label), due to a ceiling effect. We amended this in a subsequent study, by using the same design structure but with more complex stimuli. Precisely, we increased the dimensionality of our stimuli, from two to three features, keeping one salient non-diagnostic head, and two non-salient features, a tail and feet. Importantly, neither the tail nor the feet were reliably diagnostic of category membership, and as such, participants had to pay attention to both those non-salient features. As before, all participants were provided with non-linguistic feedback to learn the categories, and we expected participants who

heard a redundant label to learn the categories more quickly and to look more at the diagnostic features. On the contrary, we found that participants in the label condition looked significantly more at the non-diagnostic head at the beginning of training, while their looking behaviour at the end of training did not differ from that of participants in the no-label condition. Further, participants in the two groups did not differ with respect to the number of exemplar presentations they needed to correctly learn the categories, another result that did not replicate previous findings. We however explained these results by considering jointly the counter-intuitive nature of our stimuli and categories and participants' background knowledge. On the one hand, animal heads are arguably very diagnostic features in the real world, to the point that participants may have implicitly learned to look more at the head of a named creature. On the other hand, our stimuli were expressly made so that the head would not be diagnostic of category membership, forcing participants to focus on less salient features. Thus, participants in the label condition first looked more at the head, usually diagnostic. This may have led to an impairment in their ability to learn the categories at first, explaining why we did not see the expected facilitatory effect of redundant labels on learning categories in adults.

Importantly, the main limitation of our two adult studies was that we used the same or similar stimuli as we did for 15-month-old infants, in an attempt to allow for a comparison of the results between adults and infants to better understand the possible developmental differences between early and mastered category learning. Although this is often seen in the literature, the comparison is usually done between older children and adults, rather than young infants (e.g. Deng & Sloutsky, 2012, in which 4- to 5-year-old preschool children and adults were tested). This in turns allows for (a) the use of stimuli of appropriate complexity for both age groups, and (b) the use of the exact same design and empirical measures across age groups. This second point is particularly noteworthy here: there was a discrepancy in the design of our studies and the empirical measures used between adults and infants, which makes the comparison between the two harder to defend. Infants were presented with an implicit categorisation task, without any knowledge of the number or nature of categories, whereas adults were explicitly instructed to sort stimuli into two categories. Furthermore, infants in the control condition had no incentive to form categories and we have no way of knowing if and which categories they formed; we can only guess that infants in the control group might have formed categories based on the salient head, but it is entirely possible that they formed different categories (based on the tail or a combination of both the head and tail) or even that they grouped all the stimuli together under the same global category due to the shared body shape and colour. Conversely, adults in the control group were given the same categorisation tasks, the only difference being that they only heard non-linguistic feedback regarding their categorisation decision but no additional linguistic feedback. Crucially, adults in both conditions were given an explicit categorisation task; this, alongside testimonies from some participants, leads us to believe that adults used a conscious strategy to complete the task, most likely interfering with the unconscious effect of auditory labels we aimed to study. Therefore, a possible effect of condition

would likely be very different between our infants and adults, due to our design. It is however possible to link the two together, as in both studies, the specific effect of adding an category-defining auditory label was tested, thus we could assume that we are testing a common underlying mechanism, but testing its effect on different tasks.

Finally, our empirical results left open the question of how exactly participants, either adults or infants, learned the category, and only showed that they did learn the categories and that labels failed to attract their attention to the non-salient yet diagnostic features. To gain insights into how feature salience and auditory labels interact when learning categories, we set out to develop a computational model of categorisation that included an explicit attention mechanism. First, we chose to re-use the neurocomputational structure implementing the *labels-as-features* theory that we used earlier in this thesis to successfully replicate empirical data from 10-month-old infants. Then, we proposed an implementation of attention based on the generally accepted idea that a higher attention in general terms (for example longer looking, or investing more cognitive resources into a look) will result in better learning, and reciprocally, a lower level of attention will result in less learning. Thus, we added to our model an ‘attention bias’ that reduced the learning rate for low-salient features. We then reproduced the experimental design we used on infants, in which labels corresponded to the non-salient feature, and studied the impact of labelling depending on how differently salient the two features were. Specifically, we looked at the network’s performance in terms of learning, as measured by network error, and in terms of the compactness of the categories in its hidden representations. What we found was that, when the non-salient feature was much less salient compared to the salient feature, the model successfully encoded the different exemplars only when presented with a label. Moreover, without a label, encoding was mostly impaired for the low-salient feature, but it also impacted encoding of the highly salient feature to some extent. This result showed that the addition of a label could improve the encoding of information in general and for the diagnostic feature specifically. Importantly, we did not implement any implicit attention driving mechanisms from the label, and as such, the mere statistical covariation between the label and the diagnostic low-salience feature was enough to explain results typically associated with the *labels-as-symbols* theory. Further, for large differences in salience, the compactness of categories in the model’s hidden representations was improved by the addition of a label, another sign that adding a label helped the model build strong category representations.

Overall then, this thesis provides a first step into testing directly possible attentional effects of labelling on categorisation by controlling for the salience of object features in empirical studies and computational modelling, and predicts through computational modelling that labels will have a different impact on categorisation depending on how different the salience of multiple object features is.

Although our empirical data on infants indirectly suggested that they treated labels as features, further work is needed to confirm this finding and rule out possible confounding effects. For example, it is possible that, in our design, the head was too salient compared to the tail, and that running a similar experiment with a smaller difference

in diagnosticity would lead to different results, as suggested by our modelling work. It is also possible that, in the same way that adding a second non-salient diagnostic feature substantially changed the looking pattern results in adults, increasing the category complexity will have an impact on infants' looking patterns.

One particular question of interest would be to confirm the mechanisms by which auditory labels drove adults' attention to the salient head in our design, and to further study when these mechanisms arise in development. If, as we hypothesised, this looking behaviour results from adults' background knowledge that heads are usually diagnostic, pinpointing when in development this link emerges, and how it emerges, would shed a new light on the role of auditory labels in categorisation. In particular, further work is needed to test whether labels actively drive attention to features that are known to be diagnostic in a top-down way, as suggested by the *labels-as-symbols* theory, or if this increased attention to typically-diagnostic features is merely due to a statistical association between labels and those features as suggested by our modelling work.

Another important point that needs further studying is in clarifying the amount and nature of information that eye-tacking data give us, and finding other measures to possibly complement it. Indeed, our work on infants added to the growing evidence that learning, as evidenced clearly at test, can be seen without any statistical differences in terms of looking behaviours during training. One candidate that has been put forward to delve deeper into attention mechanisms is pupil dilation, even if the underlying physiological phenomena are still debated.

Crucially however, most of our statistical analyses resulted in non-significant results. As such, they did not provide evidence that the different groups behaved similarly, but merely failed to provide evidence that they behaved differently. Specifically, the 'Sample Theory Based' (STB) statistics that we used do not differentiate between inconclusive evidence and evidence in favour of a true null effect. Bayesian statistics provide us with tools to more finely analyse the evidence that the data provides in favour of either the null or an alternative hypothesis, and as such, their use would greatly enrich our results. However, Bayesian analysis tools have mostly been documented, in psychology, for adult data, and the guidelines developed for these data do not allow for an effective use of Bayesian analysis on infant data. Infant data is indeed notoriously noisy compared to adult data.

The first Bayesian analysis tool we meant to use on our data were Bayes factors obtained from the comparison of nested models of increasing complexity. This follows the STB approach described in Chapter 2, replacing the p -values obtained through a likelihood ratio test by b -values obtained through bridge sampling of the models' posterior distributions and drawing a Bayes factor from those distributions. The b -values thus obtained represent the likelihood, given the data at hand, of one model over the other model with one parameter removed. Although this approach has been widely spread over the last decade (e.g. Dienes, 2014; Wetzels et al., 2011), due to its straightforward use as direct alternative to p -values, it has some drawbacks (Kruschke, 2013). One typical issue is that b -values are often analysed on their own, when all they

do is inform us on the updated likelihood of some model given new data, when this should be put in perspective with the initial likelihood of the model: if a model was very unlikely to be true, but was given a very high b -value given a dataset, it would still be very unlikely true, only less so unlikely. However, even accepting that this tool is not flawless, its use on infant data is difficult, as a substantially higher number of participants is required to achieve similar power for the very lenient criterion of a b -value greater than three, as compared to finding a p -value lesser than 0,05 (roughly one-and-a-half to twice as many participants to reach 80% power according to some recent unpublished simulations available online¹). This explains why, even with an up-to-standards sample size of 48 participants, model comparisons led to mostly inconclusive b -values.

The second Bayesian analysis tool we considered was the use of a *region of practical equivalence* (ROPE) introduced by Kruschke (2013). In this framework, a small interval around the null value is defined for each model parameter, and is used to make a judgement based on each parameter's posterior distribution. As such, parameters whose posterior distribution fall entirely within the ROPE are deemed to follow the null distribution for practical purposes, meaning that a true difference for the null would anyway be too small to bear any importance on a practical point of view. On the opposite, parameters whose posterior distribution fall entirely outside the ROPE are viewed as truly different from the null distribution, and finally no conclusion can be drawn for parameters whose distributions only partly overlap with the ROPE. When studying infants however, the noisiness of the data is such that it is impossible to define a ROPE that would be wide enough to encompass true null effects, and yet narrow enough to allow for the detection of substantial effects.

Finally, we could not use Bayesian analysis to get more information from our data, and more work is needed to develop tools that would allow for the use of Bayesian analysis on infant data. For example, it is known that the specification of priors can influence the computation of Bayes factors; one can thus imagine that better understanding what priors are appropriate when studying infant data might lead to more meaningful b -values. It is also possible that other tools will be developed, or that Bayesian analysis for infant data will be rendered possible by the emergence of a wider availability of open data that could then be brought together and analysed within a Bayesian framework.

Concluding Remarks This thesis answered a few questions, but importantly, it raised many more new questions to be answered.

The stimuli and category structure we used in our empirical work were designed to be simple enough for 15-month-old infants to succeed in an implicit categorisation task. We however aimed to compare those infants with adults, using the exact same stimuli, on an explicit categorisation task, resulting in a ceiling effect on adult performance. One way to address this issue while still comparing expert adult learners and children still developing their categorisation abilities would be to test older children (4- to 5-year-old), on more complex stimuli, and crucially using identical methods for both adults and

¹<https://github.com/respatte/mb4-analysis/blob/master/Simulations.Rmd>

children (e.g. explicit or implicit categorisation). It would also be easier to design an experiment meaningful for both young infants and children, in order to gain a better understanding of the whole developmental trajectory.

Notably, our attempt to run an adult study with more complex stimuli provides additional insight on how to design future stimuli. Namely, having only two features that were not fully diagnostic did not prove difficult enough, and the complexity needs to be increased further. Previous studies have successfully used humanoid five-featured stimuli (head, torso, feet, hands, antennae) on both children and adults, future studies could therefore adapt our category structure concept to such stimuli. Further, the high success rate in learning the categories in infants suggests that they might also be able to successfully process more complex stimuli and category structures in such lab-based implicit categorisation tasks. Thus, future studies could increase the complexity of our infant task, and could expect to find different looking behaviours as infants have to engage more cognitive resources to make sense of the label-category link.

Our studies yielded many non-significant results, and the “sample theory based” framework we used did not allow us to ascertain whether or not our data supported the null hypothesis or simply did not provide conclusive evidence. The use of Bayesian statistics should allow us to better understand the evidence provided by our data, but these statistics have yet to be adapted for infant studies. Future studies would benefit from Bayesian statistics better framed for infant studies and the naturally noisy infant data, in terms of default or informative priors to use, tools and criteria to test for the importance of an effect, and more generally guidelines on how to run and interpret Bayesian statistics in infant studies. These guidelines might be better ascertain on large-scale studies such as conducted by the ManyBabies consortium.

Finally, if our modelling work provided insights on the possible links between observed behaviour and internal knowledge representations, it however failed to directly replicate our results. The modelling of attentional focus is however important to better understand how attention can affect learning, and more work is needed to determine how best to model this at a conceptual levels on auto-encoders, which have been and are still being used to model looking times in infants and adults, tasks in which visual attention plays an important role. In parallel to this, more work is needed to understand precisely what information eye-tracking data convey, and how other tools such as EEG or pupillometry might complement it. Taken together, this will help us better understand how computational models relate to empirical measures, and in turns how those empirical measures relate to cognitive processes and theories.

Bibliography

- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Rain-cloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, 4, 63. <https://doi.org/10.12688/wellcomeopenres.15191.1>
- Althaus, N., & Mareschal, D. (2012). Using saliency maps to separate competing processes in infant visual cognition: Separating competing processes. *Child Development*, 83(4), 1122–1128. <https://doi.org/10.1111/j.1467-8624.2012.01766.x>
- Althaus, N., & Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Transactions on Autonomous Mental Development*, 5(4), 288–297. <https://doi.org/10.1109/TAMD.2013.2264858>
- Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PloS one*, 9(7), e99670. <https://doi.org/10.1371/journal.pone.0099670>
- Althaus, N., & Plunkett, K. (2015a). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, 1–11. <https://doi.org/10.1111/desc.12358>
- Althaus, N., & Plunkett, K. (2015b). Timing matters: The impact of label synchrony on infant categorisation. *Cognition*, 139, 1–9. <https://doi.org/10.1016/j.cognition.2015.02.004>
- Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*, 151, 5–17. <https://doi.org/10.1016/j.jecp.2015.11.013>
- Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1), 48–53. <https://doi.org/10.1111/j.1467-7687.2007.00563.x>
- Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of experimental child psychology*, 64(1), 3–26. <https://doi.org/10.1006/jecp.1996.2332>
- Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development*, 60(2), 381. <https://doi.org/10.2307/1130984>
- Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Effects of linguistic labels on visual attention in children and young adults. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.00358>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). <https://doi.org/10.18637/jss.v067.i01>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119. <https://doi.org/10.1016/j.jecp.2013.05.002>
- Birnholz, J., & Benacerraf, B. (1983). The development of human fetal hearing. *Science*, *222*(4623), 516–518. <https://doi.org/10.1126/science.6623091>
- Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, *81*(3), 884–897. <https://doi.org/10.1111/j.1467-8624.2010.01440.x>
- Bronson, G. W. (1991). Infant differences in rate of visual encoding. *Child Development*, *62*(1), 44. <https://doi.org/10.2307/1130703>
- Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2018). Neurocomputational models capture the effect of learned labels on infants' object and category representations. *IEEE Transactions on Cognitive and Developmental Systems*, 1–1. <https://doi.org/10.1109/TCDS.2018.2882920>
- Capelier-Mourguy, A., Twomey, K. E., & Westermann, G. (2019). Learning categories with conflicting feature salience and diagnosticity. *PsyArXiv*.
- Caporello Bluvas, E., & Gentner, T. Q. (2013). Attention to natural auditory signals. *Hearing Research*, *305*, 10–18. <https://doi.org/10.1016/j.heares.2013.08.007>
- Charlesworth, W. R. (1969). The role of surprise in cognitive development. In *Studies in cognitive development. essays in honor of jean piaget* (D. Elkind & J. Flavell (Eds.), pp. 257–394). Oxford, England, Oxford University Press.
- Chen, Y.-C., & Westermann, G. (2012). Twelve-month-old infants learn crossmodal associations between visual objects and natural sounds in ecologically valid situations. *Seeing and Perceiving*, *25*(0), 117. <https://doi.org/10.1163/187847612X647504>
- Cohen, L. B. (1973). A TWO PROCESS MODEL OF INFANT VISUAL ATTENTION. *Merrill-Palmer Quarterly of Behavior and Development*, *19*(3), 157–180.
- Colombo, J., Mitchell, D. W., Coldren, J. T., & Freesean, L. J. (1991). Individual differences in infant visual attention: Are short lookers faster processors or feature processors? *Child Development*, *62*(6), 1247. <https://doi.org/10.2307/1130804>
- Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, *23*(2), 178–186. <https://doi.org/10.1177/0956797611429133>

- Deng, W., & Sloutsky, V. M. (2015). Linguistic labels, dynamic visual features, and attention in infant category learning. *Journal of Experimental Child Psychology*, *134*, 62–77. <https://doi.org/10.1016/j.jecp.2015.01.012>
- Deng, W., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24–62. <https://doi.org/10.1016/j.cogpsych.2016.09.002>
- Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dink, J., & Ferguson, B. (2018). *eyetrackingR* [R package version 0.1.8]. R package version 0.1.8.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52. <https://doi.org/10.2307/2282330>
- Edmiston, P., & Lupyan, G. (2015). What makes words special? words as unmotivated cues. *Cognition*, *143*, 93–100. <https://doi.org/10.1016/j.cognition.2015.06.008>
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107. <https://doi.org/10.1037/0096-3445.127.2.107>
- Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences*, *110*(38), 15231–15235. <https://doi.org/10.1073/pnas.1221166110>
- Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child development*, *81*(2), 472–479. <https://doi.org/10.1111/j.1467-8624.2009.01408.x>
- Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, *105*(1), 218–228. <https://doi.org/10.1016/j.cognition.2006.09.005>
- Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. <https://doi.org/10.1080/19345747.2011.618213>
- Gelman, S. A., & Coley, J. D. (1991). Language and categorization: The acquisition of natural kind terms. In *Perspectives on language and thought: Interrelations in development* (pp. 146–196). Cambridge, Cambridge University Press.
- Gelman, S. A., & Waxman, S. R. (2009). Response to sloutsky: Taking development seriously: Theories cannot emerge from associations alone. *Trends in Cognitive Sciences*, *13*(8), 332–333. <https://doi.org/10.1016/j.tics.2009.05.004>
- Gliga, T., Volcic, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience*, *22*(12), 2781–2789. <https://doi.org/10.1162/jocn.2010.21427>
- Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738. <https://doi.org/10.1111/j.1551-6709.2009.01026.x>

- Graham, S. A., & Diesendruck, G. (2010). Fifteen-month-old infants attend to shape over other perceptual properties in an induction task. *Cognitive Development*, 25(2), 111–123. <https://doi.org/10.1016/j.cogdev.2009.06.002>
- Graham, S. A., & Poulin-Dubois, D. (1999). Infants' reliance on shape to generalize novel labels to animate and inanimate objects. *Journal of child language*, 26(2), 295–320. <https://doi.org/10.1017/S0305000999003815>
- Hendrickson, K., Walenski, M., Friend, M., & Love, T. (2015). The organization of words and environmental sounds in memory. *Neuropsychologia*, 69, 67–76. <https://doi.org/10.1016/j.neuropsychologia.2015.01.035>
- Heyes, C. (2017). When does social learning become cultural learning? *Developmental Science*, 20(2), e12350. <https://doi.org/10.1111/desc.12350>
- Hilton, M., Twomey, K. E., & Westermann, G. (2019). Taking their eye off the ball: How shyness affects children's attention during word learning. *Journal of Experimental Child Psychology*, 183, 134–145. <https://doi.org/10.1016/j.jeep.2019.01.023>
- Hilton, M., & Westermann, G. (2017). The effect of shyness on children's formation and retention of novel word–object mappings. *Journal of Child Language*, 44(6), 1394–1412. <https://doi.org/10.1017/S030500091600057X>
- Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, 2. <https://doi.org/10.3389/fpsyg.2011.00017>
- Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, 13(4), 341–348. <https://doi.org/10.1002/icd.364>
- Hurley, K. B., & Oakes, L. M. (2015). Experience and distribution of attention: Pet exposure and infants' scanning of animal images. *Journal of Cognition and Development*, 16(1), 11–30. <https://doi.org/10.1080/15248372.2013.833922>
- Jankowski, J. J., Rose, S. A., & Feldman, J. F. (2001). Modifying the distribution of attention in infants. *Child Development*, 72(2), 339–351. <https://doi.org/10.1111/1467-8624.00282>
- Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517), 1–10. <https://doi.org/10.1080/01621459.2016.1240079>
- Jusczyk, P. W. (2000). *The discovery of spoken language* (1st MIT Press). Cambridge, MA, MIT Press.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1480. <https://doi.org/10.1109/5.58325>
- Kovic, V., Plunkett, K., & Westermann, G. (2009). Eye-tracking study of animate objects. *Psihologija*, 42(3), 307–327. <https://doi.org/10.2298/PSI0903307K>
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22–44. <https://doi.org/10.1037/0033-295X.99.1.22>

- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. <https://doi.org/10.1037/a0029146>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
- Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means* [R package version 1.3.5.1]. R package version 1.3.5.1.
- Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & Frank, M. (2016). A quantitative synthesis of early language acquisition using meta-analysis. *PsyArXiv*. <https://doi.org/10.17605/osf.io/htsjm>
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. <https://doi.org/10.1037/0033-295X.111.2.309>
- Lüdtke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.
- Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348–369. <https://doi.org/10.1037/0096-3445.137.2.348>
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. <https://doi.org/10.1111/j.1467-9280.2007.02028.x>
- Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception & Psychophysics*, *72*(8), 2236–2253. <https://doi.org/10.3758/APP.72.8.2236>
- Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186. <https://doi.org/10.1037/a0024904>
- Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, *8*(3), 291–318. [https://doi.org/10.1016/S0885-2014\(93\)80003-C](https://doi.org/10.1016/S0885-2014(93)80003-C)
- Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition*, *59*(3), 307–335. [https://doi.org/10.1016/0010-0277\(95\)00696-6](https://doi.org/10.1016/0010-0277(95)00696-6)
- Mani, N., Durrant, S., & Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. *Journal of Memory and Language*, *66*(4), 612–622. <https://doi.org/10.1016/j.jml.2012.03.003>
- Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science*, *21*(7), 908–913. <https://doi.org/10.1177/0956797610373371>
- Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition*, *121*(2), 196–206. <https://doi.org/10.1016/j.cognition.2011.06.013>
- Mareschal, D., & French, R. M. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06

- Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology, 36*(5), 635–645. <https://doi.org/10.1037//0012-1649.36.5.635>
- Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review, 117*(1), 1–31. <https://doi.org/10.1037/a0018130>
- Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(1), 37–50. <https://doi.org/10.1037/0278-7393.8.1.37>
- Merzenich, M. M., Van Vleet, T. M., & Nahum, M. (2014). Brain plasticity-based therapeutics [Publisher: Frontiers]. *Frontiers in Human Neuroscience, 8*. <https://doi.org/10.3389/fnhum.2014.00385>
- Mirolli, M., & Parisi, D. (2005). Language as an aid to categorization: A neural network model of early language acquisition. In *Modeling language, cognition and action* (pp. 97–106).
- Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development, 11*(3), 255–268. <https://doi.org/10.1080/15248371003699977>
- Oakes, L. M., Coppage, D. J., & Dingel, A. (1997). By land or by sea: The role of perceptual similarity in infants' categorization of animals. *Developmental psychology, 33*(3), 396.
- Oakes, L. M., Madole, K. L., & Cohen, L. B. (1991). Infants' object examining: Habituation and categorization. *Cognitive Development, 6*(4), 377–392. [https://doi.org/10.1016/0885-2014\(91\)90045-F](https://doi.org/10.1016/0885-2014(91)90045-F)
- Oakes, L. M., & Plumert, J. M. (2002). Variability in thirteen-month-old infants' touching patterns in the sequential-touching task. *Infant Behavior and Development, 25*(4), 529–549. [https://doi.org/10.1016/S0163-6383\(02\)00149-2](https://doi.org/10.1016/S0163-6383(02)00149-2)
- Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language, 135*, 66–72. <https://doi.org/10.1016/j.bandl.2014.05.005>
- Perry, L. K., & Lupyan, G. (2016). Recognising a zebra from its stripes and the stripes from “zebra”: The role of verbal labels in selecting category relevant information. *Language, Cognition and Neuroscience, 1*–19. <https://doi.org/10.1080/23273798.2016.1154974>
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition, 106*(2), 665–681. <https://doi.org/10.1016/j.cognition.2007.04.003>
- Quinn, P. C. (2004). Development of subordinate-level categorization in 3-to 7-month-old infants. *Child Development, 75*(3), 886–899. <https://doi.org/10.1111/j.1467-8624.2004.00712.x>

- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Rakison, D. H., & Butterworth, G. E. (1998). Infants' use of object parts in early categorization. *Developmental Psychology*, *34*(1), 49–62. <https://doi.org/10.1037/0012-1649.34.1.49>
- Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study, In *34th international conference on machine learning*. 34th International Conference on Machine Learning, Sydney, Australia.
- Rivera, S., & Robinson, C. W. (2016). Learning in the wild-how labels influence what we learn. 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA.
- Roberts, K., & Jacob, M. (1991). Linguistic versus attentional influences on nonlinguistic categorization in 15-month-old infants. *Cognitive Development*, *6*(4), 355–375. [https://doi.org/10.1016/0885-2014\(91\)90044-E](https://doi.org/10.1016/0885-2014(91)90044-E)
- Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*(5), 1387–1401. <https://doi.org/10.1111/j.1467-8624.2004.00747.x>
- Robinson, C. W., & Sloutsky, V. M. (2007a). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, *11*(3), 233–253. <https://doi.org/10.1111/j.1532-7078.2007.tb00225.x>
- Robinson, C. W., & Sloutsky, V. M. (2007b). Visual processing speed: Effects of auditory input on visual processing. *Developmental Science*, *10*(6), 734–740. <https://doi.org/10.1111/j.1467-7687.2007.00627.x>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.
- Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2003). Infant visual recognition memory: Independent contributions of speed and attention. *Developmental Psychology*, *39*(3), 563–571. <https://doi.org/10.1037/0012-1649.39.3.563>
- Roy, D. (2003). Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, *5*(2), 197–209. <https://doi.org/10.1109/TMM.2003.811618>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Salverda, A. P., & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(4), 1122–1133. <https://doi.org/10.1037/a0023101>
- Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space (J. Wiles, Ed.). *PLoS ONE*, *6*(12), e28095. <https://doi.org/10.1371/journal.pone.0028095>

- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584. <https://doi.org/10.1146/annurev.ps.46.020195.003021>
- Sloutsky, V. M. (2009). Theories about ‘theories’: Where is the explanation? comment on waxman and gelman. *Trends in Cognitive Sciences*, *13*(8), 331–332. <https://doi.org/10.1016/j.tics.2009.05.003>
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166–188. <https://doi.org/10.1037/0096-3445.133.2.166>
- Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, *111*(1), 65–86. <https://doi.org/10.1016/j.jecp.2011.07.007>
- Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child development*, 1695–1709. <https://doi.org/10.1111/1467-8624.00373>
- Sloutsky, V. M., & Napolitano, A. C. (2003). Is a picture worth a thousand words? preference for auditory modality in young children. *Child Development*, *74*(3), 822–833. <https://doi.org/10.1111/1467-8624.00570>
- Sloutsky, V. M., & Robinson, C. W. (2008). The role of words and sounds in infants’ visual processing: From overshadowing to attentional tuning. *Cognitive Science: A Multidisciplinary Journal*, *32*(2), 342–365. <https://doi.org/10.1080/03640210701863495>
- Träuble, B., & Pauen, S. (2007). The role of functional information for infant categorization. *Cognition*, *105*(2), 362–379. <https://doi.org/10.1016/j.cognition.2006.10.003>
- Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*, 413–438. <https://doi.org/10.1111/cogs.12539>
- Twomey, K. E., & Westermann, G. (2015). A neural network model of curiosity-driven infant categorization, In *Development and learning and epigenetic robotics (ICDL-EpiRob), 2015 joint IEEE international conference on*, IEEE. <https://doi.org/10.1109/DEVLRN.2015.7346097>
- Twomey, K. E., & Westermann, G. (2017a). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, e12629. <https://doi.org/10.1111/desc.12629>
- Twomey, K. E., & Westermann, G. (2017b). Learned labels shape pre-speech infants’ object representations. *Infancy*, *23*, 61–73. <https://doi.org/10.1111/inf.12201>
- Waxman, S. R., & Booth, A. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, *6*(2), 128–135. <https://doi.org/10.1111/1467-7687.00262>

- Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. <https://doi.org/10.1016/j.tics.2009.03.006>
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Westermann, G., & Mareschal, D. (2004). From parts to wholes: Mechanisms of development in infant visual object processing. *Infancy*, *5*(2), 131–151. https://doi.org/10.1207/s15327078in0502_2
- Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, *27*(4), 367–382. <https://doi.org/10.1016/j.cogdev.2012.08.004>
- Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391–20120391. <https://doi.org/10.1098/rstb.2012.0391>
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*(3), 291–298. <https://doi.org/10.1177/1745691611406923>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
- Yamauchi, T., Kohn, N., & Yu, N.-Y. (2007). Tracking mouse movement in feature inference: Category labels are different from feature labels. *Memory & Cognition*, *35*(5), 852–863. <https://doi.org/10.3758/BF03193460>
- Yamauchi, T., & Yu, N. -Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition*, *36*(3), 544–553. <https://doi.org/10.3758/MC.36.3.544>
- Younger, B. A., & Furrer, S. D. (2003). A comparison of visual familiarization and object-examining measures of categorization in 9-month-old infants. *Infancy*, *4*(3), 327–348.