Representational and sampling assumptions drive individual differences in single category generalisation

Keith Ransom (keith.ransom@adelaide.edu.au)
School of Psychology, University of Adelaide
Andrew T. Hendrickson (a.hendrickson@tilburguniversity.edu)
Cognitive Science & Artificial Intelligence, Tilburg University
Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne
Danielle J. Navarro (d.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

Human activity requires an ability to generalise beyond the available evidence, but when examples are limited – as they nearly always are – the problem of how to do so becomes particularly acute. In addressing this problem, Shepard (1987) established the importance of *representation*, and subsequent work explored how representations shift as new data is observed. A different strand of work extending the Bayesian framework of Tenenbaum and Griffiths (2001) established the importance of *sampling assumptions* in generalisation as well. Here we present evidence to suggest that these two issues should be considered jointly. We report two experiments which reveal replicable qualitative patterns of individual differences in the representation of a single category, while also showing that sampling assumptions interact with these to drive generalisation. Our results demonstrate that how people shift their category representation depends upon their sampling assumptions, and that these representational shifts drive much of the observed learning.

Keywords: categorisation; generalisation; representations; sampling assumptions;

Introduction

Suppose that, upon encountering a wallaby for the first time, I am reliably informed that *wallabies are dax*. What should I infer to be the extension of the property *dax*? If I know that *dax* is a biological property I might generalise to other macropods, marsupials, or mammals. Alternatively, if *dax* describes a behaviour I might instead generalise to other hopping or grazing animals. As this thought experiment suggests, human category representations are structured and complex; multiple systems of categories are relevant to a single domain and different systems of knowledge are relevant in different contexts (Heit & Rubinstein, 1994; Ross & Murphy, 1999).

Although there is some work investigating how people acquire multiple systems of categories (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011) and learn which representations are relevant to inductive problems like this (Austerweil & Griffiths, 2010), very little is known about individual differences in representation. Do such differences exist, and can they be measured? When people learn based on new data, do their representations shift? If so, how and why? Do their assumptions about how the data were generated drive any of this? These are the questions we focus on in this paper.

Representation and generalisation

The problem we consider is ostensibly a simple one: learning how to generalise along a single stimulus continuous dimension. Stimulus generalisation in this situation often resembles an exponential decay as a function of distance along the relevant dimension, but only when formulated with respect to the proper stimulus representation (Shepard, 1987). When adapting Shepard's analysis into an explicitly Bayesian framework, Tenenbaum and Griffiths (2001) noted that generalisation from multiple examples allows for many different possible stimulus representations. Indeed, there are many different assumptions a learner might make about category representation. These include exemplar models (Nosofsky, 1986), prototype models (Smith & Minda, 1998), decision boundaries (Ashby & Townsend, 1986), critical regions that mimic prototype models if the regions are connected (Tenenbaum & Griffiths, 2001), or exemplar models in which each item corresponds to a region (Navarro, 2006). Additionally, these representations are not fixed and stable, Evidence from category learning has shown that human learners tend to "grow" category representations as they see additional items, with a shift from prototype to exemplar representations during learning (Griffiths, Canini, Sanborn, & Navarro, 2007; Love, Medin, & Gureckis, 2004) or a mixture of representations across individuals (Kalish & Kruschke, 1997).

Sampling and generalisation

An adjacent literature on inductive generalisation has revealed that what the learner assumes about how this data came to be the data has a substantial influence on the inferences people draw. These *sampling assumptions* affect inferences in concept learning tasks (Navarro, Dry, & Lee, 2012), property induction tasks (Ransom, Perfors, & Navarro, 2016), and word learning problems (Xu & Tenenbaum, 2007).

While there are many possible sampling assumptions that one might adopt (e.g. Shafto, Goodman, & Griffiths, 2014; Ransom, Voorspoels, Perfors, & Navarro, 2017), much of the literature has focused on two simple possibilities. A helpful teacher is likely to choose positive examples that belong to the relevant category (known as strong sampling), whereas a random sampling process selects exemplars independently of the category label (known as weak sampling). The dif-

ference between the two leads to a variety of differences in how people generalise: most notably, people tend to *tighten* their generalisations with additional data if they are assuming strong sampling, but don't if they aren't (e.g., Xu & Tenenbaum, 2007; Ransom et al., 2016).

Sampling and representation?

If both representation and sampling assumptions shape generalisation, how do they fit together? The literature on sampling assumptions typically assumes a fixed stimulus representation, and the literature on stimulus representation has given little consideration to the manner in which exemplars are chosen. In this paper, we present empirical evidence suggesting that these two problems should be considered together. We report results from two experiments involving a simple inductive generalisation task that manipulates the sampling assumptions across conditions. We find evidence for individual differences in category representation, with different participants appearing to represent categories in different ways. Moreover, there appears to be an interaction between people's representations and the degree to which they are sensitive to the sampling manipulation. Observations selected by a helpful teacher are more likely to cause people to *shift* their mental representation of the category in a consistent direction than if the same observations are selected at random. In fact, these representational shifts seem to account for the largest share of learning in the task.

Experiment 1

Experiment 1 is a single category generalisation experiment that, within the same experimental framework, combines manipulations of sample size (as in Navarro et al., 2012; Vong, Hendrickson, Perfors, & Navarro, 2013) and sampling cover story (as in Ransom et al., 2017; Xu & Tenenbaum, 2007). As a post-hoc analysis, we use people's responses across all test items to identify clusters of people who generate similar patterns of generalisation. These patterns are then used as predicted outcomes in Experiment 2, where they are explicitly connected to representational clusters. Furthermore, the assignment of individual behaviour to clusters is tracked during learning, in order to determine whether representational shifts correspond to learning outcomes.

Method

Participants 603 people participated in this experiment via Amazon Mechanical Turk, where they were paid $1.30US for the 5-10 minute task. 45% were female, 93% were from the US, and median age was 32 (range: 19 to 77).

Design People were randomly assigned to one of three conditions that varied the number of category exemplars ("Wuggams") as well as the manner in which they were sampled. In the FOUR condition ($N = 194$) participants were shown four exemplars with no explanation offered for how these examples were chosen. Participants in the TWELVE HELPFUL ($N = 200$) and TWELVE RANDOM ($N = 209$) conditions were also shown the same four exemplars with no explanation, but were then subsequently shown eight more exemplars for which an explanation was given. In the TWELVE HELPFUL condition people were told that the additional examples had been intentionally chosen to help them understand the category, whereas people in the TWELVE RANDOM condition randomly selected additional items themselves.

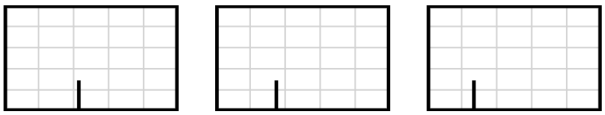


Figure 1: **Example stimuli**. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

Stimuli Stimuli consisted of a black rectangular frame drawn against a white background, with a vertical black line inside attached to the bottom edge (see Figure 1). To assist with stimulus discriminability, four evenly spaced light grey vertical and horizontal lines were drawn within the rectangle. Stimuli varied along a single dimension, corresponding to the horizontal position of the vertical line within the rectangle (referred to later as the *stimulus value*).

The full set of training stimuli included 12 examples with stimulus values ranging from 21% to 43% in increments of 2%. People in the TWELVE HELPFUL and TWELVE RANDOM conditions saw all 12 examples, while those in the FOUR condition saw four, including the two extreme examples (at 21% and 43%) plus two random others in between. The test stimuli consisted of 19 items with values ranging from 5% to 95% in increments of 5%.

Procedure The experiment consisted of a training phase where people were shown examples from the target category, followed by a test phase where they were asked to decide whether previously unseen items were in that category.

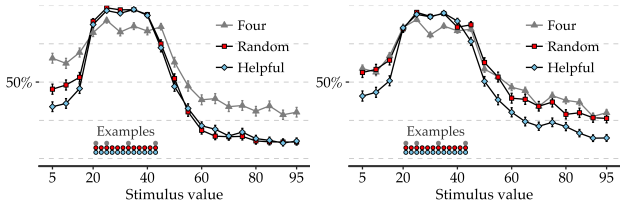
Training. Participants were told that the purpose of the experiment was to see how people judged whether or not unfamiliar objects were in the same category as known examples. In the FOUR condition the instructions stated:

> So, we'll start by showing you some objects that all belong to the same category («Wuggams»).

at which point four training examples were displayed simultaneously on-screen. Participants in the the other two conditions were given the same introduction. However, after the initial examples were shown those in the TWELVE RANDOM condition were further informed:

> **The computer has assigned you to experiment group «J8»** so we're going to let you pick an additional «8» items at random from our collection, and let you see any «Wuggams» that you find.

Following this a 6×5 arrangement of icons resembling packing boxes was displayed on screen, and people were asked to select eight boxes one by one. After clicking on an icon the image was replaced with that of an open box, people



(a) Experiment 1. (b) Experiment 2.

Figure 2: Performance on a one category generalisation task as a function of sampling procedure (manipulated between subjects) and sample size (manipulated between subjects in Experiment 1 and within subjects in Experiment 2). The graphs show the proportion of positive responses to the question: "Do you think this object is in the «Wuggam» category?" for each of the test stimuli. The performance of people who saw four examples of the target category (grey line) is contrasted with two groups of people who saw 12 examples (black lines). In Experiment 1, people tightened their generalisations as more data is observed, but the sampling manipulation had little effect; whether people actively sampled the additional examples at random (red squares) or were told that the items had been selected by a helpful teacher (blue diamonds), they generalised less when they saw 12 examples rather than 4. In Experiment 2, where the wording of the sampling manipulation was slightly adjusted, tightening with increased sample size occurs, but only in the HELPFUL condition.

were informed that they had found a «Wuggam» inside, and one of the training examples was added to the display. The TWELVE HELPFUL condition proceeded along similar lines, but people were instead told:

> **The computer has assigned you to experiment group «K8»** so we're going to help you by showing you an additional «8» «Wuggams» **chosen by a helpful teacher** to give you a good idea of the full range of «Wuggams».

Following this, the array of boxes was displayed with eight of the boxes already opened. Simultaneously, the display was updated with the eight additional examples. In all conditions the on-screen presentation order was randomised.

Testing. To minimise any memory effects, the training examples remained on screen during testing, along with a reminder of how the exemplars were selected. Participants in all conditions were shown the 19 test stimuli one at a time in random order; this sequence was repeated four times. The test query was a simple yes or no question, "Do you think this object is in the «Wuggam» category?"

Results and discussion

The results are shown in Figure 2(a), which plots the proportion of trials on which each test item was assigned to the Wuggam category in each condition. There is a clear effect of sample size: people who saw 12 examples generalised to a narrower range of test items than those who saw 4. A Bayesian ANOVA reveals strong evidence ($BF_{10} > 10^6$) for a model that includes effects of stimulus value, sample size and an interaction, tested against a null model that includes only the effect of stimulus value.[1] However, the cover story appeared to have little to no effect, with modest evidence favouring the null hypothesis ($BF_{01} = 10$) that generalisation patterns were the same in both 12-item conditions.

The one exception to this pattern is the three test items to the far left of Figure 2(a). Visual inspection suggests that participants in the TWELVE HELPFUL condition were somewhat less willing to generalize to these items than were people in the TWELVE RANDOM condition. This asymmetric pattern is not predicted by "standard" implementations of the Bayesian generalisation model (e.g. Navarro et al., 2012; Vong et al., 2013). However, it is consistent with a shift in the proportion of people using a single decision boundary, which should not fall off on the far (left) side of the observed exemplars.

To examine this possibility we conducted a post hoc clustering analysis of generalisation curves at the individual subject level. This analysis, which was based on a Dirichlet process mixture model, automatically identified 11 different "patterns" of generalisation curves. Nine of the 11 patterns accounted for 98% of the data; and of these nine, three were minor variants of the others.[2] The remaining six patterns, which form the core of the analysis in Experiment 2, cover 85% of the data from that experiment. We turn to it next.

Experiment 2

Participants 404 people participated in this experiment via Amazon Mechanical Turk, where they were paid $1.50US for the 10-15 minute task. 48% were female, 94% were from the US, and median age was 32 (range: 18 to 71).

Design, stimuli & procedure Experiment 2 was a preregistered[3] replication and extension of Experiment 1. The two experiments were identical except for three key differences. First, we adopted a within subject manipulation of sample size. Regardless of condition, participants were shown four exemplars with no sampling explanation given and then tested. They were then shown an additional eight exemplars – either within a HELPFUL (N=205) cover story or a RANDOM one (N=199) – and then tested a second time. Testing each person twice allows us to assess how their representation changed based on four examples or twelve.

Second, at the end of each test phase participants were asked to identify the strategy they used, selecting one of the six options listed in Figure 3(b). This data is useful for determining whether their reported strategies correspond to the generalisation patterns our model assigns to them.

Third, the cover story in the RANDOM condition was al-

[1] Model comparisons included a random intercept for each subject, and fit using default priors (Rouder, Morey, Speckman, & Province, 2012; Liang, Paulo, Molina, Clyde, & Berger, 2012) from the BayesFactor package (version 0.9.12-2) in R (version 3.4.3).

[2] We used the `BayesianGaussianMixture` class from the `scikit.learn` module v0.19.1) under Python 3.6.3. The concentration parameter for the Dirichlet process was set to 1, the multivariate Gaussian distribution assumed a diagonal covariance structure, and the random seed was set to 1. Each generalisation pattern was encoded as a point in 19-dimensional space with each dimension corresponding to a stimulus value included in the test items and the value along each dimension corresponding to the probability of generalising the category label to that test stimulus. Supplemental materials describing details of the model and all 11 patterns are here: `https://tinyurl.com/RPNH18`

[3] `https://aspredicted.org/3tq89.pdf`

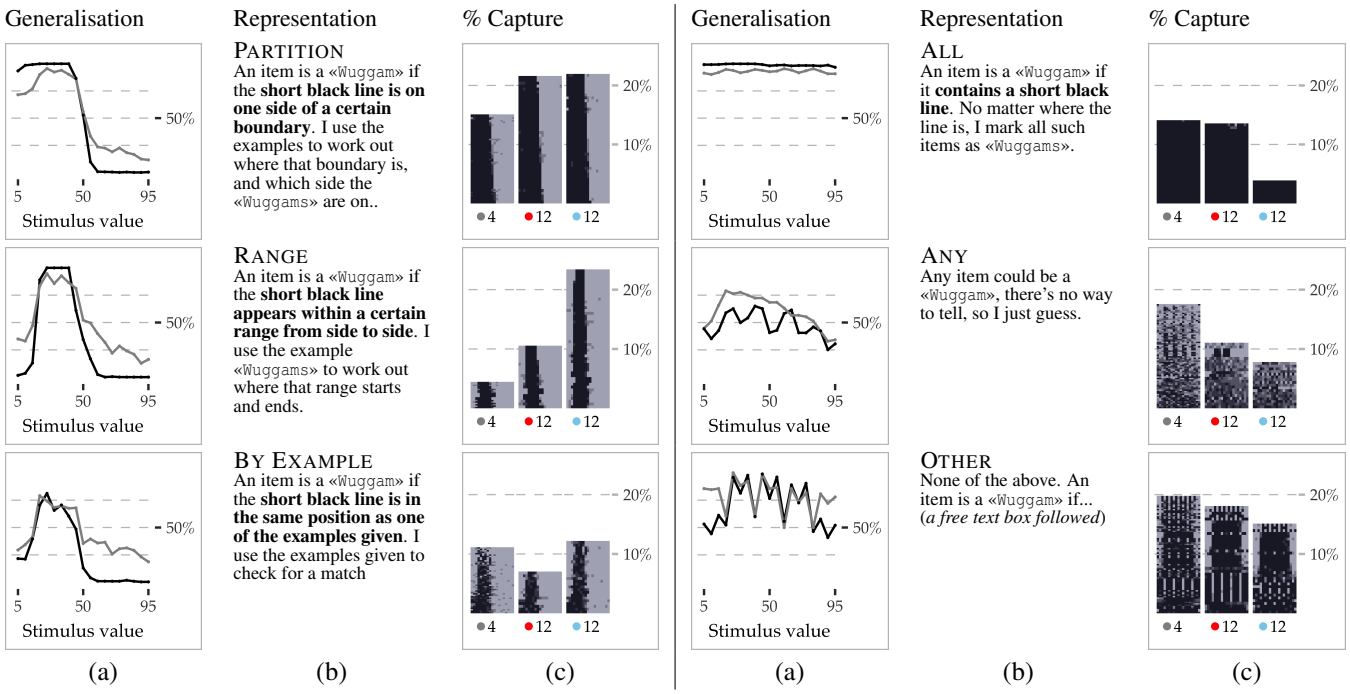


Figure 3: A graphical depiction of individual differences in generalisation in Experiment 2. The panel columns represent: (a) Aggregate generalisation curves for people grouped by data driven pattern definition (black lines) and by response to self report question (grey lines). (b) The response options for the questions that asked people about their response strategy (title added). There is a one-to-one mapping between the patterns shown and the representation associated with each response option. (c) The proportion of people allocated to a given pattern. The three bars from left to right represent people after seeing four examples, and after seeing 12 examples in the RANDOM (red) and HELPFUL (blue) conditions respectively. The rows of pixels within each bar constitutes a grey-scale representation of the generalisation data of individuals in that pattern and condition (see main text for detail). Both sample size and sampling assumption impact people's representation of the target category.

tered slightly in order to leave open the possibility that some boxes might not contain Wuggams. People were told that "some of the boxes are stuck and won't open; in that case just try another." Each person sampled 11 boxes but saw only 8 «Wuggams» in total; the other three times (when the box remained closed) occurred in a random order with the constraint that the first and last item was always a «Wuggam».

Results and discussion

Generalisation Generalisation patterns in Experiment 2 partially replicated the results from Experiment 1, as shown in Figure 2(b). As before, we find a clear effect of sample size ($BF_{10} > 10^6$), but unlike Experiment 1 we also find an effect of the sampling manipulation. On an aggregate level, people in the HELPFUL condition tightened their generalisations ($BF_{10} > 10^6$) whereas those in the RANDOM condition did not ($BF_{01} = 31$). This suggests that the changed wording in the RANDOM condition, which provided a mechanism for potentially seeing a non-«Wuggam», helped to make the sampling cover story believable.

Representational analysis Our primary question was whether people used different representations and whether their representations shifted in different conditions or with extra data. To address this, we used the six main generalisation patterns identified in Experiment 1, shown in Figure 3(a). They are each suggestive of qualitatively different mental representations: a one-sided decision boundary (Partition), a two-sided Range, several different kinds of non-contiguous regions (By Example, Any, Other), and an assignment of All test items to the category. Each participant at each test phase in Experiment 2 was then separately assigned to the most similar pattern using the model derived from the results of Experiment 1.

The results of this analysis are displayed in Figure 3. First, we note that the six patterns identified by our model are indeed roughly equivalent to the six self-report options offered during the test phase (shown in the middle panels (b)).[4] This is clear when we compare the black lines in panel (a) on the left (which plot the average response for all people assigned to the relevant pattern) to the grey lines in the same panels (which plot the average generalisation curve for all people who chose the relevant self-report option). In most respects, the grey and black curves mirror each other very closely, illustrating that the data-derived patterns (based on classifications) and self-reported strategy are very similar.

[4] Alignment of the self-report to the model-identified patterns was done based on our qualitative assignment, but we also performed all analyses using assignments based on RMSE fit (which differ from the qualitative assignments for 2 of the 11 clusters), or using the (somewhat noisy) self-report data directly. In all cases the conclusions are the same. Even collapsing Partition and Range into a single representation and the remaining representations into another produces a qualitatively similar pattern of results.

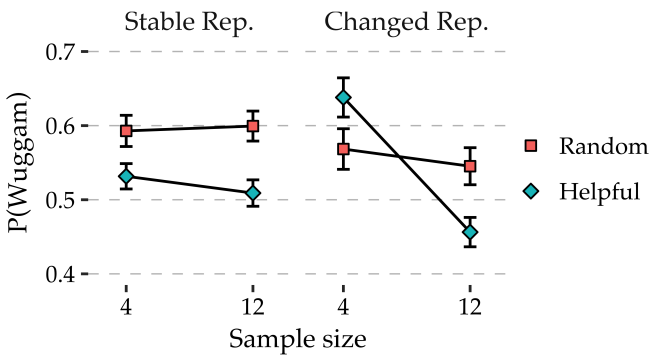


Figure 4: The mean effect of additional examples on the marginal probability of generalising the learned category to novel stimuli, as a function of sampling assumption and representational shift. Over half of the participants in Experiment 2 (N=119, from the RANDOM condition and N=111, from the HELPFUL condition) maintained a stable representation of the underlying category in response to observing an additional 8 examples, and showed little change in generalisation overall. Likewise, for people in the RANDOM condition who did undergo a representational shift (N=80). But for many people in the HELPFUL condition (N=94), the additional examples led to a representational shift resulting in a significant and consistent contraction in generalisation overall.

Although the six patterns shown in Figure 3 are quite dissimilar to one another, there is a remarkable degree of within-pattern homogeneity, especially with respect to the first four patterns: most people assigned to a pattern do genuinely appear to be closely matching that pattern. This can be seen in Figure 3(c), which depicts a compressed grayscale representation of the raw responses for every participant within a pattern. Each panel shows three bars corresponding to one of the three possible conditions (4 exemplars, 12 exemplars RANDOM and 12 exemplars HELPFUL). The height of each bar captures how many people's generalisations best matched that pattern (thus, for instance, many more people matched the Range pattern in the HELPFUL condition than any other). Within each bar, every row of black pixels displays the responses of a single participant: each row consists of 19 cells, each colour coded to represent the probability of assigning the relevant test item to the «Wuggam» category. For instance, an all black row occurs if all items are assigned to the «Wuggam» category, whereas a grey bar with a patch of black in the middle would represent a generalisation pattern where only the middle group of test items were labelled as «Wuggams».

Representational shifts We are now in a position to address the central questions motivating this experiment. To what extent are changes in generalisation driven by a change in people's *representation* of the underlying category structure (e.g., shifting from Partition to Range), as opposed to learning the parameters of a representation (e.g., learning where the boundary in a partition lies)? Do sampling assumptions have an effect on how people shift their representations?

To investigate this, note that the rightmost panels of Figure 3 are effectively bar charts, displaying the proportion of participants assigned to each of the six patterns, broken down by experimental condition. Visual inspection reveals marked differences as a function of sample size: when only 4 exemplars are observed, people are most likely to be assigned to the Any or Other patterns, whereas by the time 12 exemplars are observed the generalisation patterns are closer to Partition, Range or Other. Similarly there is evidence of a sampling effect: helpful sampling guides learners towards a Range representation whereas random sampling does not. A Bayesian contingency test ($BF_{10} > 10^6$) finds strong evidence for a difference in pattern assignments across conditions.

Looking more closely at these data, we can examine whether the sampling conditions each had a different impact on how people shifted their representations. To do so, we used the representation label assigned to each generalisation pattern (see Figure 3). If people were assigned to one pattern after seeing four examples and a different pattern after seeing 12 examples, *and* those patterns had different representation labels, then and only then would they be considered as having undergone a representational shift. Figure 4 plots the results of this analysis. It is clear that additional exemplars led to an overall narrowing of generalisation, but only for those people who believed the examples were selected by a helpful teacher, and largely as a consequence of a representational shift. A Bayesian ANOVA reveals strong evidence ($BF_{10} > 10^6$) in favour of a model that includes effects of sample size, sampling condition, representational shift and interactions, tested against a null model which includes only the participant as a random nuisance parameter.

General discussion

The present work examines how people generalise a concept on the basis of learned examples. In a single experimental framework, we jointly considered two important considerations known to shape such generalisation: namely, people's assumptions about how the data was sampled, and their representation of the concept they seek to generalise.

In an initial between-subject experiment we found an effect of sample size consistent with other inductive generalisation tasks of this kind (e.g Navarro et al., 2012; Vong et al., 2013). While there was no aggregate effect of sampling assumption, a post hoc analysis of individual responses revealed common patterns of generalisation suggestive of mental representations explored in the literature. These included non-contiguous regions (Nosofsky, 1986), a one-sided decision boundary (Ashby & Townsend, 1986), and a two-sided connected region (Tenenbaum & Griffiths, 2001). This analysis also suggested that people's sampling assumptions might play a role in determining their representation of the category, a hypothesis we tested in a second pre-registered experiment.

The second experiment was based closely on the first but was within-subjects and involved a random sampling cover story that was slightly modified to be more suggestive of weak sampling. It replicated the effect of sample size and also found an effect of the revised sampling manipulation. Moreover, by linking response patterns identified in the first exper-

iment to people's responses in the second, we found that observing additional examples causes some people to undergo a change in their mental representation. This shift drove much of their change in generalisation, and the nature and consistency of the change critically depended upon people's sampling assumptions.

In many ways our results are consistent with previous work finding that people tighten their generalisations when strong sampling holds but fail to do so when it does not (Xu & Tenenbaum, 2007; Ransom et al., 2016; Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015). This work has attributed such tightening to the operation of the size principle, which favours smaller hypotheses in a fixed (researcher defined) hypothesis space (Tenenbaum & Griffiths, 2001). However, our results suggest that while the size principle may still be at work in some fashion, the truth may be more complex. Learning may in fact be operating on at (at least) two levels in an hierarchical space, one with different representations (hypothesis spaces) at the top level and fixed hypotheses within each representation at the lower level. Behaviour that on aggregate looks like generalisation according to the size principle may actually reflect individuals shifting their representations more than individuals tightening their generalisations within the same representational space. An interesting line for future research would be to attempt to account for this behaviour using a hierarchical model that learns on both of these levels.

One open question is to what extent our results rely on the patterns identified by the Dirichlet process Gaussian mixture model. Although we are reasonably reassured by the close qualitative convergence between its assignments and people's responses, another possibility would be to investigate other methods for automatically learning mental representations. For instance, recent advances in deep neural network (DNN) architectures has spiked considerable interest in the relationship between the representations used by humans and those learned by DNNs. In preliminary work we employed two known DNN architectures as a means of dimensional reduction: the first method used a stacked auto-encoder with layer-wise pre-training and a sparsity penalty, while the second used a variational auto-encoder. Both techniques, when fit to the data from the first experiment, yielded a clustering solution for the second experiment that was comparable or better to the one reported here. We plan to explore the utility of these approaches in future work.

Conclusion

"Is this a dagger which I see before me...?" – Macbeth
"That's not a knife. That's a knife." – Crocodile Dundee

On the question of how best to classify sharp pointy things, great literary protagonists differ. And life, in this respect, may imitate art. Individual differences in representations, known to be driven by data, may be driven by sampling assumptions as well. By taking such differences seriously we have begun to understand that relationship; we hope that further work in this direction will continue to yield richer insights.

Acknowledgments

This work was supported by an Australian Government Research Training Program Scholarship (KR) and ARC Discovery Grant DP180103600.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.
Austerweil, J., & Griffiths, T. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 32).
Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process.
Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 411.
Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1362–1377.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 410–423.
Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309–332.
Navarro, D. J. (2006). From natural kinds to complex categories.
Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.
Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796.
Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. (2017). A cognitive analysis of deception without lying. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.* (pp. 992–997).
Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495–553.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.
Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.
Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(04), 629–640.
Vong, W. K., Hendrickson, A. T., Perfors, A., & Navarro, D. J. (2013). The role of sampling assumptions in generalization with multiple categories. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3699–3704). Austin, TX: Cognitive Science Society.
Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? nonmonotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1-25.
Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.