Exploring the role that encoding and retrieval play in sampling effects

Keith Ransom (keith.ransom@adelaide.edu.au)
School of Psychology, University of Adelaide
Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Abstract

A growing body of literature suggests that making different sampling assumptions about how data are generated can lead to qualitatively different patterns of inference based on that data. However, relatively little is known about how sampling assumptions are represented or when they are incorporated. We report the results of a single category generalisation experiment aimed at exploring these issues. By systematically varying both the sampling cover story and whether it is given *before* or *after* the training stimuli we are able to determine whether encoding or retrieval issues drive the impact of sampling assumptions. We find that the sampling cover story affects generalisation when it is presented before the training stimuli, but not after, which we interpret in favour of an encoding account.

Keywords: categorisation; generalisation; memory; sampling assumptions;

Introduction

For most of the reasoning tasks with which we are routinely faced, it is impossible to draw conclusions that are logically entailed by what we know already. Instead, we must by necessity make inductive generalisations on the basis of the limited data we have. In order to make the most of that data, it is important to accurately assess its evidentiary weight – to recognise precisely what kind of generalisations it supports. Doing this assessment accurately depends on understanding the context in which it was observed.

To illustrate why, imagine that you need to buy a present for a colleague as a part of your workplace Secret Santa. You don't know this colleague that well, but while helping them move offices you see a box containing the CDs that they listen to while at work. Sensing an opportunity to re-gift an unwanted copy of *Taylor Swift*, you take a closer look. Upon realising that almost all of their collection consists of 80s Billboard Hits, you conclude that their musical taste is dated[1] and reluctantly decide that Taylor Swift is not for them.

Suppose, instead, that you had seen the exact same data (a box of CDs) but in the context of helping your colleague move their entire music collection – many dozens of boxes worth – and that box just happened to be the only open one. Now the same data is no longer quite so representative: instead of being a carefully culled and chosen set of favourites, it is one of many. Thus, it tells you much less about whether your colleague would like Taylor Swift.

As this example illustrates, knowing something about why one saw the data that one did (and not some other data) enables people to make more valid inferences. Put another way, being able to reason about the generative process behind a set of observations tells people about the weight of evidence that those observations supply. These assumptions about the generative process are often referred to as the *sampling assumptions* that people bring to inference problems. Different sampling assumptions appear to drive qualitatively distinct patterns of generalisation (e.g. Hendrickson, Perfors, Navarro, & Ransom, 2019; Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019), support epistemic trust (Shafto, Eaves, Navarro, & Perfors, 2012) and epistemic vigilance (Landrum, Eaves, & Shafto, 2015; Ransom, Voorspoels, Perfors, & Navarro, 2017), fuel pragmatic implicature (Goodman & Frank, 2016), and promote accelerated learning (Shafto, Goodman, & Griffiths, 2014).

Despite this wealth of empirical support for the utility and importance of sampling assumptions in generalisation, little is known about either how they affect the encoding and retrieval of the data, or how they affect people's mental representations. Is the evidentiary weight of data under a given sampling assumption computed only at the point at which the data is later retrieved? Or is it encoded at the time of learning, thus shaping the underlying representation from the beginning? And how is inference affected as people's memories of the data begin to fade?

Using a single-category learning task, we explore these questions here for the first time. We manipulate both the sampling assumptions people make about the training data (via cover story) as well whether that cover story is available before or after learning. As we explain in the next section, if sampling assumptions affect generalisation at retrieval, we expect no difference in performance regardless of when the cover story was revealed. Conversely, if they affect how the data are encoded, we expect different patterns of generalisation depending on when the cover story was available.

Sampling assumptions and inductive generalisation

The Bayesian generalisation approach of Tenenbaum and Griffiths (2001) provides a useful framework for our research question. In the context of our single category generalisation experiment, we are interested in how the learner decides whether or not to extend the target category c to a novel item y on the basis of previously observed examples x. Within the framework, this decision is assumed to be probabilistic, based on the available evidence. That is:

$$P(y \in c|x,s) = \sum_{h \in \mathcal{H}_c : y \in h} P(h|x,s) \quad (1)$$

where s represents the learner's assumption about the process generating the data x, and $\mathcal{H}_c$ represents the set of alternative hypotheses the learner considers concerning the true extent

[1] The fact that your colleague still uses CDs may have told you this already.

of the category c.[2] In other words, the evidence in favour of category membership is effectively combined across all hypothetical versions of the category containing the novel item. Using a straightforward application of Bayes' rule the term $P(h|x,s)$ may be expressed as:

$$P(h|x,s) \propto P(x|h,s)P(h). \quad (2)$$

This formulation assumes, for simplicity, that the learner entertains a single sampling assumption (i.e. $P(s) = 1$), which we presume was given to them by a cover story describing the generative process.

It is the likelihood function $P(x|h,s)$ that is critical for our current purposes. Substituting different likelihood functions into this system of equations yields different predictions about the way that people generalise from given data. For instance, strong sampling implies a likelihood that embodies the size principle, such that each subsequent datapoint serves as evidence to further tighten one's generalisations around the data; weak sampling uses a different likelihood which implies no such tightening (Tenenbaum & Griffiths, 2001). Thus, the likelihood may be thought of as representing different ways of calculating the weight of evidence that the data provides for the hypothesis under a given sampling assumption.

Our first question here is *when* the likelihood is calculated: when the data is first encoded, or when it is retrieved? If learners do not need to rely on their memories and the sampling cover story is available from the beginning, it is impossible to disentangle these two possibilities. However, if we manipulate when participants are aware of how the data were sampled (i.e., before or after learning), then different possibilities yield different predictions. We consider two main possibilities in detail.

Retrieval. If the likelihood is calculated upon retrieval, then encoding need only involve storing the raw data x in some form. The likelihood calculation would be shaped by whatever sampling assumption was in play during retrieval, regardless of what was assumed during learning. In this sense, the calculation would resemble the conventional or "idealised" interpretation of the Bayesian generalisation model. However, while the conventional interpretation assumes perfect recall of exemplars, a failure to retrieve some data would imply that the likelihood calculation was effectively over a reduced dataset (i.e., smaller sample size). The precise effect that this has will depend on the sampling assumption and on the particular items forgotten. For example, if the diversity of the dataset is largely unaffected by the failure to retrieve certain items, then generalisation under a strong sampling assumption should be wider in this case than under perfect recall. Under weak sampling, in contrast, it is the diversity of the sample and not its size that has an effect on generalisation; thus, a reduction in sample size without a change in diversity would mean that generalisation was unaffected. More generally, as the level of retrieval failure increases, the Bayesian model predicts generalisation increasingly in line with the prior distribution.

[2] In the case that the data x varies over a continuous dimension $\mathcal{H}_c$ will represent a continuum of hypotheses and the sum is replaced with an integral.

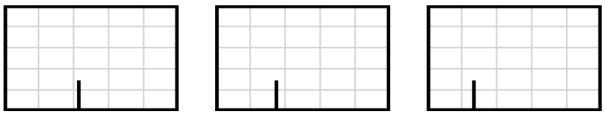


Figure 1: **Example stimuli**. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

Encoding. If the likelihood is calculated upon encoding, then the strength of evidence that it represents would have to be stored in some way. In this case, the precise effect of later retrieval failure might vary depending on *how* evidence is encoded. For example, if evidence is stored and retrieved with each exemplar individually then failure to retrieve a given exemplar would mean that subsequent generalisation operates over a smaller dataset, as in the retrieval account (although, unlike the retrieval account, using the sampling assumption that was in play at the time of encoding). If instead, evidence were stored and retrieved in aggregate form (via the hypotheses, for example) then failure to recall any particular exemplar need not imply that the associated evidence was lost. In this way, generalisation might still proceed with all the available evidence (presuming the same hypotheses were accessed). The details of representation notwithstanding, if the likelihood is calculated and stored during encoding, and not at retrieval, then generalisation would be shaped by the sampling assumptions available during learning, even if those assumptions are changed at retrieval.

Method

Our experiment involved a single-category generalisation task modelled on previous work demonstrating that sample size and sampling cover story affect people's willingness to extend category membership to novel examples (Hendrickson et al., 2019; Ransom, Hendrickson, Perfors, & Navarro, 2018). Although we employed stimuli identical to those used in that experiment, we modified the method of presentation so that each stimulus was removed from screen after a (typically brief) period of self-paced study. Using a consistent experimental framework allows us to directly compare our results with the previous findings, and thus to determine if the effect of sampling assumptions on generalisation changes as the memory of training examples decays.

One of our manipulations involved the nature of the cover story people received. Either they were told that the data was given by a HELPFUL teacher (which corresponds to a strong sampling assumption and implies that generalisations should be tighter) or they were given a cover story implying that it was chosen at RANDOM (which corresponds to a weak sampling assumption and implies that generalisations should be looser). Critically, we manipulated whether people were given the sampling story BEFORE or AFTER they saw the training stimuli. If sampling assumptions affect how

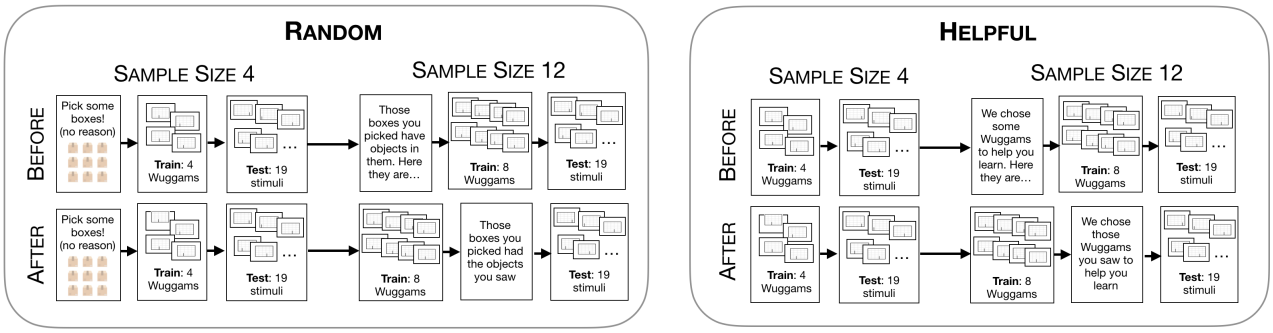


Figure 2: **Experiment design**. Our 2x2x2 design varied Sample Size within-subject and Sampling Explanation and Presentation Sequence between-subjects. All participants began by seeing four individually-presented exemplars followed by a generalisation task to novel stimuli. Those in the RANDOM condition were then given a cover story in which the subsequent eight items were chosen at random from boxes that they themselves had previously selected. Those in the HELPFUL condition were told that the items were selected by a helpful teacher. In the BEFORE condition, the cover story was given before seeing the eight new items; in the AFTER, it came after. In all conditions the experiment ended with a repeat of the generalisation test.

the data are encoded then people should generalise differently depending on when they received the story.

Participants

We recruited 999 people via Amazon Mechanical Turk who were each paid $1.70USD for 5-10 minutes participation. 56% were female, with age varying between 18 and 75 (median: 37 years), drawn predominately from the U.S. population (99%). All participants passed a screening for English language competency prior to participation.

Stimuli

Stimuli were black rectangles containing a vertical black line inside, attached to the bottom edge (see Figure 1). They varied along a single dimension (the *stimulus value*): the horizontal position of the line within the rectangle. Participants were told that this was the way in which stimuli varied. Evenly spaced light grey "guide lines" were drawn within each rectangle in order to improve discriminability. There were 12 training stimuli in total, whose stimulus values ranged from 21% to 43% in increments of 2%. They were divided into two sets corresponding to the two training phases, as described below.

Design and procedure

As shown in Figure 2, our experiment employed a $2 \times 2 \times 2$ mixed factorial design. Two factors (Sampling Explanation and Presentation Sequence) were manipulated between-subjects while another (Sample Size) varied within-subject. People were thus allocated at random to one of four experimental groups.

Across all groups, the experiment involved presenting people with a number of examples of a novel 1D category and then observing whether they generalised category membership to new items based on the examples they had been shown and what they had been told about those examples.

Sample Size To facilitate a baseline against which the effect of additional exemplars could be compared, the experiment involved two rounds of testing. The first (Size 4) occurred after a training phase involving four training examples, and the second (Size 12) after seeing eight more.

Stimuli for the first training phase consisted of the two extreme examples (with values of 21% and 43%) and two others selected at random from the ten whose values lay between the extremes. The eight remaining stimuli formed the second training set and were presented in random order.

Presentation Sequence This between-subjects manipulation varied when the sampling cover story was presented in relation to the second training set. People in the BEFORE condition were told the cover story (RANDOM or HELPFUL, described below) *before* viewing the second set of training items, while people in the AFTER condition were offered the explanation only after all training items had been presented.

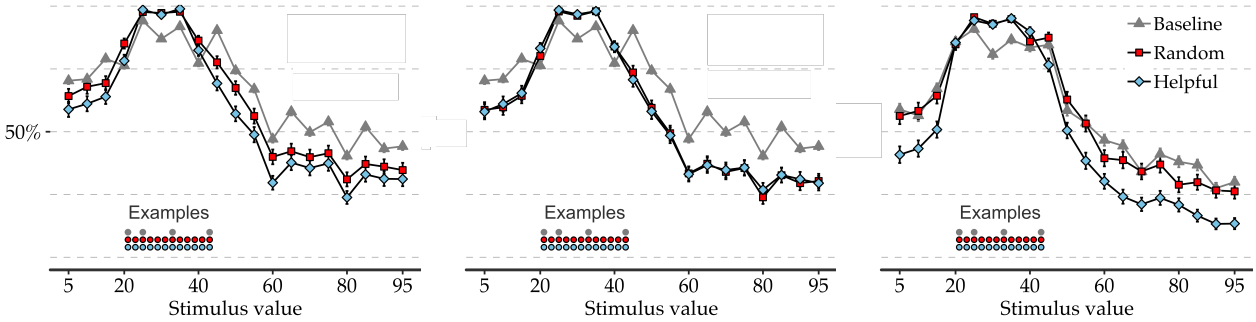
Sampling Explanation The other between-subjects manipulation varied the details of the cover story explaining how the data in the second training phase were generated. The initial training phase, however, was identical for all participants. No explanation was given for how the exemplars were chosen. People were told only 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 second training phase people were given one of two different cover stories explaining how the items were selected.

Helpful. People in the HELPFUL condition were told:

> We have a bunch of boxes containing examples of the full variety of «`Wuggams`». We have chosen 8 of these boxes especially to help you learn the «`Wuggam`» category, bearing in mind the four training examples we showed you originally.

at which point an array of eight icons resembling open packing boxes were displayed in an adjacent panel. Participants in the BEFORE condition then viewed the eight stimuli one by one. Those in the AFTER condition saw the identical explanation (with verb tenses adjusted) only after all eight stimuli in the second training phase had been shown.

Random. The RANDOM condition was designed to encourage people to believe that each training item was selected



(a) BEFORE. (b) AFTER. (c) Ransom et al. (2018).

Figure 3: Performance on a one category generalisation task as a function of presentation sequence, sampling procedure (manipulated between-subjects) and sample size (manipulated within-subject). 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. People's performance after seeing four examples of the target category with no sampling explanation given (grey line) is contrasted with their performance after seeing all 12 examples and being given an explanation of how the additional examples were selected (black lines). (a) When the sampling explanation was given prior to the presentation of the final 8 examples (BEFORE condition), people tightened their generalisations as more data was observed, but the extent of tightening was affected by the sampling manipulation; those people who actively sampled the additional examples at random (red squares) tightened their generalisation less than those that were told that the items had been selected by a helpful teacher (blue diamonds). (b) In contrast, when the sampling explanation was given only after all training stimuli were presented (AFTER condition), the sampling manipulation had no effect, with people tightening their generalisation equally in both cases. (c) Using the same experimental framework and stimuli, but keeping the training stimuli on-screen during the testing phase, Ransom et al. (2018) demonstrated the effect of sampling manipulation seen only in the BEFORE condition. But when people must rely on their memory of observed examples, their generalisation is wider overall.

at random and that it was at least theoretically possible to see examples not in the target category. To achieve this, people in the RANDOM condition were presented with an additional phase preliminary to the first training round. In this phase, a 6×5 arrangement of packing boxes was displayed on screen, and people were asked to select boxes in any order (but not told why this was necessary). After selecting 11 boxes, people were told that the contents would be revealed later in the experiment. Following this, the first training phase commenced, which was identical for all participants.

During the second training phase, participants in the AFTER condition were immediately shown the eight remaining training items without explanation. Those in the BEFORE condition were told that we had many boxes containing examples from our catalogue, and that these examples included but were not limited to Wuggams. After this, the original array of (closed) boxes was displayed, indicating the ones that the participant had previously selected. People were then told:

> At the start of the experiment we asked you to choose some of these boxes at random. These are the boxes that you selected. We're going to open them now and show you whatever kind of item we find inside.

In order to reinforce the notion that it might have been possible to see items from categories other than Wuggams, the display was updated at this point to reveal eight open boxes and three closed ones. People were told that some of the boxes they had chosen were stuck but that we would show them the contents of the boxes that did open. Participants in the AFTER condition received exactly this cover story (with verb tenses adjusted) only after seeing all eight training examples.

Generalisation test

Immediately after both the first and second training phase, participants in all conditions performed the same generalisation test. In it, they were shown 19 stimuli one at a time in random order; this sequence was repeated four times. The stimuli consisted of 19 items with stimulus values ranging from 5% to 95% in increments of 5%. The test query was a yes or no question: "Do you think this object is in the «Wuggam» category?" Neither training stimuli nor the sampling explanation remained on-screen during testing, requiring people to rely on their memory when making judgements.

Results

Our work is focused on understanding how memory and sampling assumptions interact to affect generalisation. Do we replicate previous findings showing that differences in sampling assumptions lead to differences in generalisation? Does this difference in people's patterns of generalisation change if the sampling manipulation occurs before or after stimulus encoding? We address each question in turn below.

First: do we replicate previous results? Our RANDOM BEFORE and HELPFUL BEFORE conditions are very similar to that of a previous study (Ransom et al., 2018), but are different in one key way. In our version, the training stimuli were removed from the screen after initial presentation; in Ransom et al. (2018) and much of this literature the training stimuli stay visible for the entire experiment. We therefore investigate whether these previously observed effects of sampling manipulation are replicated even when people must rely on their memory of the training stimuli.

To investigate this we first analysed the responses of all participants having seen only the first four exemplars, for which no sampling explanation was given. Against this baseline we separately compared the responses of people in the RANDOM BEFORE and HELPFUL BEFORE conditions. The resulting generalisation curves shown in Figure 3(a) reveal

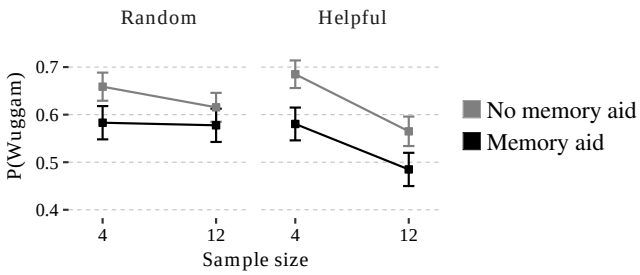
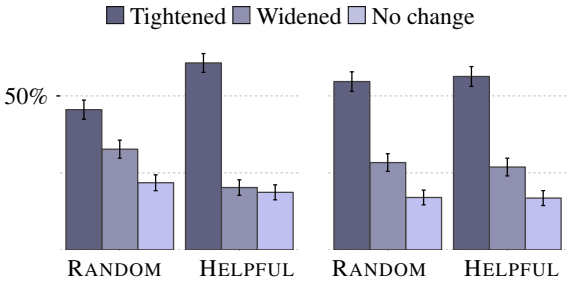


Figure 4: The mean effect of additional exemplars on the marginal probability of generalising the learned category to novel stimuli, as a function of sampling assumption and the presence of a memory aid. When training exemplars remained on-screen throughout the testing phase participants were less willing overall to generalise the target category to novel items than when no memory aid was present. In magnitude, the effect of the memory aid on generalisation was comparable to the effect of observing the eight additional exemplars.



(a) BEFORE (b) AFTER

Figure 5: The proportion of people who either tightened ($\Delta_p < 0$), widened ($\Delta_p > 0$) or showed no change ($\Delta_p = 0$) in their region of generalisation, after seeing additional examples (where Δ_p reflects an individual's change in rates of responding in favour of the learned category). People are grouped according to the explanation they received about the sampling of extra items, and whether it was given before or after the examples themselves. Error bars show standard error of proportion. (a) In the BEFORE condition, where the sampling explanation was given prior to the presentation of the additional examples, the sampling manipulation had an effect. The majority of people who were told that the items had been selected by a helpful teacher tightened their region of generalisation, while the (slight) majority of people in the RANDOM condition, who actively sampled their own additional examples, widened their region of generalisation or showed no change. (b) In contrast, when the sampling explanation was provided after the additional stimuli had been presented (as in the AFTER condition), the majority of people tightened their generalisations regardless of the explanation given.

that the HELPFUL sampling manipulation led to tighter generalisation than the RANDOM manipulation. This replicates a key finding of Ransom et al. (2018), shown in Figure 3(c). To examine the strength of evidence for this finding we analysed generalisation curves for the second test phase (Size 12), calculating the generalisation probability for each person and stimulus separately. A Bayesian ANOVA revealed that a model of generalisation probability including stimulus value and sampling manipulation as predictors is strongly preferred to a model containing stimulus value only ($BF_{10} > 10^6$).

Although we replicated the qualitative difference between sampling conditions, it is evident on visual comparison of Figure 3(a) and (c) that people appeared to generalise further when they had to rely on their memory of the training stimuli. To determine the overall effect that this had on generalisation we calculated the marginal probability of extending category membership to novel items as a function of test phase (4 or 12 items) and sampling manipulation (RANDOM or HELPFUL). We then compared this probability between our experiment (the BEFORE conditions) and Ransom et al. (2018).

The results, shown in Figure 4, demonstrate that the absence of a memory aid had a uniform but significant effect on generalisation overall ($BF_{10} > 10^{100}$).[3] After seeing 12 exemplars, participants in our study (who had no memory aid) showed a willingness to generalise to novel items comparable to participants in Ransom et al. (2018) after seeing only four items that remained on screen throughout. Thus, overall, we find that the *difference* in generalisation according to sampling assumption did replicate, but generalisation was consistently higher when people had to rely on their memory more.

Our second question was whether the effect of sampling manipulation changes when the sampling cover story is given after the training stimuli rather than before. We therefore repeated our analysis for people in the RANDOM AFTER and HELPFUL AFTER conditions, and found that it does: there is no longer a difference in generalisation based on sampling assumption. As Figure 3(b) shows, people tighten their generalisations to a remarkably similar degree across the two conditions, despite the fact that they had opposing sampling cover stories (Bayesian ANOVA now favours the model with stimulus value as the only predictor: $BF_{01} = 42$).

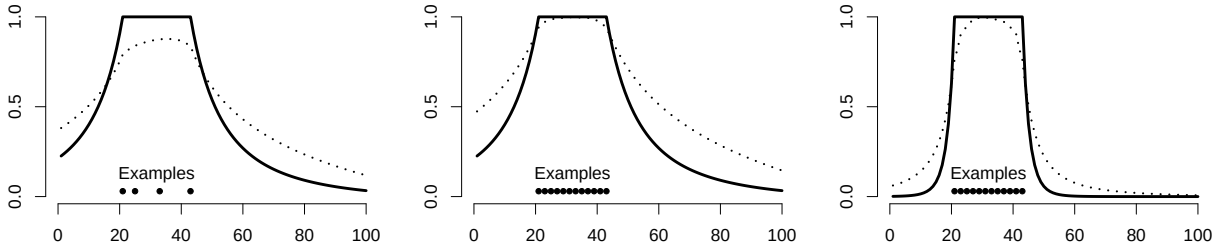
To further assess the effect of our sampling manipulation on the qualitative patterns of responding, we compared each individual's responses between the two test phases, after seeing 4 and 12 exemplars. Figure 5 shows the proportion of people who either tightened, widened or showed no net change in their generalisation (marginalised across test items). Consistent with the patterns at the aggregate level, it is evident that the explanation given to participants regarding the source of the additional exemplars does affect the trajectory of generalisation as more examples are observed. But this explanation only has an effect if it is given before the exemplars are observed ($BF_{10} = 300$) and not after ($BF_{01} = 2.8$).[4]

Discussion

To our knowledge, our work here is the first to explore *when* sampling assumptions affect generalisation, and by extension when the likelihood is calculated. Our results demonstrate that the sampling cover story only had an effect when it was made explicit prior to the presentation of the data. When it was presented at retrieval, then whatever likelihood was the default at the time of encoding (which, in this case, ap-

[3] Based on a Bayesian logistic regression comparing a model of yes/no responses that included stimulus value, sampling manipulation and memory aid as predictors to one without memory aid.

[4] Bayes' factors are based on a multinomial logistic regression comparing a model of qualitative effect (tighten, widen, no net change) with sampling manipulation as a predictor against an intercept only model.



(a) BASELINE. (b) RANDOM. (c) HELPFUL.

Figure 6: Simulated performance on a one category generalisation task as a function of exemplar recall, sampling assumption and sample size. The graphs plot the probability of generalising the learned category as a function of stimulus value. Solid lines represent generalisation performance on the assumption that all exemplars are perfectly recalled at decision time – the default assumption of the Bayesian generalisation model. Dashed lines represent generalisation performance on the basis of imperfect recall. For illustration purposes, the simulation uses an independent probability of recall for each exemplar ($p = 0.5$). Failing to recall exemplars leads to wider generalisation overall. (a) Simulated performance in the BASELINE condition (4 exemplars), assuming the default (strong) sampling. When the sample size is small, the effect of forgetting on generalisation reflects a balance of two forces: the reduction in diversity may reduce generalisation within the range spanned by the exemplars, while the reduced sample size leads to wider generalisation outside the range. (b) Simulated performance in the RANDOM condition (12 exemplars), assuming the BASELINE performance as a prior and that the 8 additional exemplars are weakly sampled. In the case of imperfect recall, the simulation predicts that the 8 additional items, although imperfectly recalled, lead to wider generalisation as a result of increased diversity. (c) Simulated performance in the HELPFUL condition (12 exemplars), assuming the BASELINE performance as a prior and that the 8 additional exemplars are strongly sampled. Under strong sampling, generalisation tightens quickly around the sampled range with each extra exemplar, thus the predicted effect of forgetting is less in this scenario.

peared to have been strong sampling) was the likelihood that shaped generalisation – even though the cover story at retrieval should have contradicted it. While we cannot altogether rule out the influence of sampling assumptions at the point of retrieval, our experiment provides evidence in favour of an encoding account. Under this account, the evidence for different hypotheses is assessed according to the sampling assumption that prevailed at the time that the data were originally presented.

This finding has a variety of interesting implications. First, it suggests that there is no such thing as a "theoryless" learner: at no point do people simply encode the raw data in a veridical fashion. Rather, from the start they are actively engaged in making sense of it for future generalisation even though there is no current need to generalise. The question remains as to how automatic this is: would people be able to inhibit the likelihood calculation if requested to remember each specific data point as precisely as possible, or if they didn't think that a generalisation task would be forthcoming?

This has implications for effective pedagogy as well. It is known that learners benefit from assuming that their teacher is selecting the most informative examples possible given the learner's current beliefs. Such reciprocal assumptions can lead to a highly leveraged form of generalisation in which concepts can quickly be acquired from minimal input (Shafto et al., 2014). Under the idealised account of pedagogical learning, people's inferences should not depend on when the sampling process becomes apparent. However, our results suggest that it is important for the teacher to make the sampling process clear as early as possible.

In a similar way our finding has implications for how people process misinformation and corrections to misinformation. Ransom et al. (2017) found, for example, that people can use truthful but limited data in their efforts to mislead others by attempting to manipulate their counterpart's sampling assumption. Our work suggests that subsequently learning that an information source was biased may not be sufficient to correct the bias. It therefore offers another explanation for the well-established finding that retracting misinformation does not eliminate its influence (Johnson & Seifert, 1994; Ecker, Lewandowsky, Swire, & Chang, 2011). If people are encoding data in such a way that it cannot be disentangled from their theory at the time, interpreting that data under a new theory may be extremely difficult.

Another interesting aspect of this work regards the role of memory. By adopting the experimental procedure of Ransom et al. (2018) but requiring participants to view the simuli one-by-one, we were able to assess how memory decay would interact with sampling assumptions in shaping generalisation. We found that people tightened their generalisations less when they had to rely on their memory more. A simulation of the generalisation task used in our experiment verified our intuition that this should be the case (see Figure 6). Our finding is consistent with previous work using complex linguistic and non-linguistic data rather than a simple one-dimensional category (Perfors, Ransom, & Navarro, 2014), which suggests that the result is reasonably robust.

Our memory manipulation (albeit across two experiments) also provides some basis to distinguish between two possible encoding accounts. One possibility is that evidence is stored and retrieved with each exemplar individually and any failure to retrieve an exemplar would mean that computation occurs over a smaller dataset. A second possibility is that evidence is stored in aggregate (across all data points) and retrieved via the hypotheses. In this case, the contribution of each exemplar would be accounted for at the point of encoding, and so the computation should proceed as if the full dataset were retrieved. The two possibilities suggest contrasting predictions.

In the first case, we would expect generalisation in the present experiment to be wider than in the previous (Ransom et al., 2018, where perfect recall was supported). In the latter case, we should expect the results of the two experiments to be broadly in line with each other. As already noted, we found that manipulating how easy it was to remember exemplars did affect generalisation in a manner consistent with some degree of recall failure. We interpret this as weak evidence favouring the "exemplar encoding" account over the "hypothesis encoding" account: the data is stored in such a way that the strength of evidence is in some way integral to the encoding of the exemplar, at least to the extent that failure to later retrieve the exemplar equates to a failure to incorporate the associated evidence. Our evidence is only weak, however, because it is not entirely clear what "forgetting" in the context of the hypothesis encoding account would amount to. Fleshing out these distinctions more and testing them more systematically is a goal for future work.

While the present experiment should be taken in the spirit of a "proof of concept", our research nonetheless suggests that memory, sampling, and generalisation are intertwined in ways that are still not fully understood. By manipulating when different information is available as well as the cognitive load during learning, it is possible to further illuminate this complex relationship.

Acknowledgements

Thanks to Simon Dennis for helpful comments regarding the initial concept behind this study. Thanks also to the anonymous reviewers for their helpful comments. This work was supported by an Australian Government Research Training Program Scholarship (KR) and ARC Discovery Grant DP180103600

References

Ecker, U., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*(18), 570–578.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *TiCS*, *20*(11), 818–829.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, 1–8.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Johnson, H., & Seifert, C. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Jn Exp Psych: LMC*(20), 1420–1436.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109 - 111.

Perfors, A., Ransom, K. J., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. In *36th Annual CogSci Conference* (pp. 2759–2764).

Ransom, K. J., Hendrickson, A., Perfors, A., & Navarro, D. J. (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In *40th Annual CogSci Conference* (pp. 930–935).

Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. J. (2017). A cognitive analysis of deception without lying. In *39th Annual CogSci Conference* (pp. 992–997).

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

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.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Beh. & Brain Sci.*, *24*(4), 629–640.