Learning is using past experiences to inform new behaviors and actions. Because all experiences are unique, learning always requires some generalization. An effective way of improving generalization is to expose learners to more variable (and thus often more representative) input. More variability tends to make initial learning more challenging, but eventually leads to more general and robust performance. This core principle has been repeatedly rediscovered and renamed in different domains (e.g., contextual diversity, desirable difficulties, variability of practice). Reviewing this basic result as it has been formulated in different domains allows us to identify key patterns, distinguish between different kinds of variability, discuss the roles of varying task-relevant versus irrelevant dimensions, and examine the effects of introducing variability at different points in training.

Variability in everyday life

As people interact with the world through time, variability is a consistent part of our lives. Indeed, in a very real way, we never have exactly the same experience twice. What kinds of variability matter for learning, why do they matter, and when in learning do they matter most?

Consider learning to serve in tennis. One learning strategy is to always practice serving from the exact same location on the court and always aim at the exact same spot. This approach would allow a learner to quickly perfect this particular serve, but this improvement may not generalize well if the learner needs to serve from or aim at different locations. An alternative learning approach is to practice serving from various locations and aiming at various spots. Here, improvement would be slower, but the effects of training would generalize far more broadly (e.g., [1,2]). Thus, when learning to serve, increasing variability may frustrate early training, but would pay off in increased generalizability of what is learned.

The same compromise between early learning and later performance exists in many domains. For example, consider an infant learning to recognize dogs for the first time. If the infant only experiences one specific dog, they may very quickly learn to recognize it, but may struggle to recognize other dogs as dogs. Conversely, exposing infants to many different dogs would prolong initial learning of the category, but would eventually lead them to form a more robust representation of what properties make something a dog (e.g., [3]).

Similarly, in the field of language acquisition, variable input has been shown to benefit learning and generalization across multiple different levels of analysis: from speech perception [4,5], to word learning [6–9], to grammar [10–12]. For instance, infants learn to differentiate between novel words that differ in the voicing of one sound (e.g., buk and puk) only when exposed to sufficient acoustic variation in pronunciation [13,14], and adults are much better at learning new words when they appear in more variable contexts (i.e., in paragraphs on different topics) [7,8].

Highlights

For the past 80 years, the relationship between variability, learning, and generalization has been studied in various domains, including motor learning, categorization, visual perception, language acquisition, and machine learning.

Learning from less variable input is often fast, but may fail to generalize to new stimuli; learning with more variable input is initially slower, but typically yields better generalization.

This basic observation has been repeatedly reformulated under different names in different fields, but with little synthesis of similarities and differences nor recognition of different types of variability.

We highlight the complementary insights made in different domains on the role of variability in learning and integrate these insights to better understand what kinds of variability matter, when do they matter, and why.
In all, although there is certainly a great deal of nuance and domain-specific details, the general phenomenon in all of these examples is essentially the same: greater variability may initially hinder learning, but typically leads to an improved ability to generalize learning to new contexts by facilitating the formation of more abstract knowledge (e.g., [4, 15–17]; Figure 1).

Discovery and rediscovery of the importance of variability

The fundamental relationship between variability and learning outcomes has been repeatedly rediscovered and renamed in the fields of categorization, visual perception, motor learning, language, inductive reasoning, formal education, and machine learning, among others (Table 1). Researchers in these different fields have often conducted very similar studies to examine the effects of variability on learning and generalization. For example, published studies in the literature have tested whether vocabulary in a second language is learned better when learners are exposed to one versus three speakers [6], whether bean bag tosses are more accurate when practicing tossing from one versus three locations [18], whether face recognition is more accurate when people are exposed to one versus four different photos of an individual [19], and whether people become more accurate in solving verbal statistical problems when trained on one versus four examples [20]. Despite large differences in the behaviors being studied, the experimental manipulations are largely identical and have produced similar results, suggesting that the underlying principles at play are the same (Figure 1). These behavioral results have been incorporated into various computational and theoretical frameworks across fields, aimed at explaining learners’ sensitivity to variability; in particular, why do learners generalize more along dimensions that exhibit high variability and generalize less along dimensions that exhibit low variability condition.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>% of correct pattern identification</td>
<td>% of time flown within the correct glide area</td>
<td>% of correct dialect categorization score</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Illustration of similar effects of variability across three domains. A similar relationship between training stimuli and generalization has been observed in the domains of visual perception, motor learning, and language processing [4, 15, 17]. Bar plots were adapted from the reported results in the original studies. (A) Posner and Keele measured the effect of variable training on the perception of visual patterns. Compared with people who saw highly distorted patterns, people who saw less distorted patterns performed best during the learning phase but were worst at categorizing novel patterns (adapted from Table 1). (B) Huet et al. measured the effect of training variability on learning to land planes in a flight simulator. People trained under constant conditions performed better during training, but people trained under variable conditions (e.g., different runway widths) performed better on the transfer test (adapted from Figure 2). (C) Clopper and Pisoni measured the effect of talker variability on the categorization of American English dialects. People exposed to a single talker in each dialect were better during training and recognized the dialect of familiar speakers more accurately, but people exposed to three talkers in each dialect were better at recognizing the dialect of new, unfamiliar speakers (adapted from Figure 2). See [4, 15, 17].
Table 1. The names given to the variability phenomenon in different research fields

<table>
<thead>
<tr>
<th>Field</th>
<th>Term</th>
<th>Main point</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorization/visual</td>
<td>The variability effect</td>
<td>‘Variability improved classification of novel items, whereas repetition improved classification of studied items’ [91]</td>
<td>[3,15,16,19,39,40,74,91–93]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Infants can generalize a given property to new members of an animal category... only when presented with multiple exemplars of a familiarized category’ [3]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Category density</td>
<td>‘Low-variability categories can be learned with fewer observations than required to learn high-variability categories with the same means. If subjects learn two equally probable categories of unequal variability, they will tend to classify more items into the high-variability category at transfer’ [92]</td>
<td>[39–41,74–76,92–96]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘A diverse category... led to lower levels of accuracy in the training task, wider generalization, and poorer item recognition’ [40]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The spacing effect</td>
<td>‘Presenting the instances in a spaced sequence resulted in more learning than presenting the instances in a massed sequence, despite the difficulty created by the spaced sequence’ [33]</td>
<td>[32–34,55,97–101]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Compared with massing, spacing enhances long-term recall, but we expected spacing to hamper induction by making the commonalities that define a concept or category less apparent’ [97]</td>
<td></td>
</tr>
<tr>
<td>Motor learning</td>
<td>Variability of practice</td>
<td>‘Practice from a variety of locations facilitated performance when the subject was transferred to a novel location than did practice from a fixed location’ [102]</td>
<td>[1,2,17,18,102–109]</td>
</tr>
<tr>
<td></td>
<td>Contextual interference effect</td>
<td>‘Activities can be proposed in a repetitive practice schedule (blocked practice)... or in random practice schedules by performing more tasks or variations of one same activity (high interference). High contextual interference, even though causing immediate limited performance, leads to superior performance on retention and transfer tests’ [27]</td>
<td>[27–31,71,110–114]</td>
</tr>
<tr>
<td></td>
<td>distribution of practice effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Phonetic/acoustic variability</td>
<td>‘Results point to positive consequences of affective variation, both in creating generalizable memory representations for words, but also in establishing phonologically precise memories for words... High affective variation has the effect of enhancing infants' perception and retention of invariant phonological detail’ [12]</td>
<td>[12,13,46,67,115–121]</td>
</tr>
<tr>
<td></td>
<td>Talker variability</td>
<td>‘Listeners trained with high variability stimulus sets generalized well to new tokens produced by a familiar talker and to novel tokens produced by an unfamiliar talker. In contrast, listeners trained with only a single talker showed little evidence of generalization to new tokens or new talkers’ [122]</td>
<td>[4–6,9,14,52,122–125]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘When infants are exposed to variable exemplars of words, their learning is focused on the consistent pieces of information - in this case, phonological information. Infants can track which cues vary consistently within and across words, and which seem to have no connection to the words they are learning. While such a manipulation appears to make the task more difficult by adding additional irrelevant information that the infant must filter out... Multitalker training can lead infants to better word learning’ [52]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contextual/semantic diversity</td>
<td>‘Subjects were better at recognizing words after encountering them in highly variable contexts, but better at inferring their meanings after experiencing them across more stable semantic contexts’ [8]</td>
<td>[7,8,10,72,126–130]</td>
</tr>
<tr>
<td></td>
<td>Intratask interference/interleaving/spacing effect</td>
<td>‘[verbal] learning is powerfully affected by the temporal distribution of study time. Spaced (vs. massed) learning of items consistently shows benefits’ [101]</td>
<td>[101,131–138]</td>
</tr>
</tbody>
</table>

(continued on next page)
low variability [21]? The same basic relationship between variability and generalization is also observed in neural networks [22,23], which has led to attempts to optimize the generalization performance of deep learning models through artificially increasing the variability of the training input through data augmentation, such as applying transformations like rotations and color changes [24] (Box 1).

Although the effects of variability on learning and generalization have been discussed in different fields for nearly 80 years, and despite the clear commonalities of perspective across domains, there has been little to no crosstalk between these fields. Instead, the exact same principle appears under different terminology in a host of fields and their insights have not been unified into a single theoretical framework (but see [25,26] for attempts to link the literature on motor and/or verbal learning to pedagogical applications). This has obscured the bigger picture and prevented the discovery of the core principles that underlie these effects. For example, in some domains, the main questions of interest have revolved nearly exclusively around generalization (e.g., does training with more/less variability produce better/worse performance with new untrained stimuli), while in other domains they revolve mostly around learning (e.g., does training with more/less variability produce better performance with the training stimuli) (Box 2). Even in domains that focus on generalization, the term ‘generalization’ has been used in several distinct
Trends in Cognitive Sciences

In reviewing the literature, we were surprised to discover that the label ‘variability’ has been used to refer to at least four different types of variability, each stemming from a different source. This is crucial to recognize, as these different ‘variability types’ may or may not have the same impact on learning and/or generalization. Across experiments and fields, low and high variability appear to be contrasted in four different ways (Figure 2): (i) numerosity (set size), such as when learning from more or fewer distinct examples; (ii) heterogeneity (differences between examples), such as when learning from a fixed number of examples that are more or less similar to one another (this similarity, in turn, can be along task-relevant or -irrelevant dimensions, as we discuss later); (iii) situational (contextual) diversity, such as when learning from the same examples under more or less variable environmental conditions that do not pertain to the examples later); (iii) situational (contextual) diversity, such as when learning from the same examples under more or less variable environmental conditions that do not pertain to the examples.

In the following sections, we integrate the findings from different disciplines, including perception, language, and motor learning, to shed light on what kind of variability matters, why it matters, and when it matters.

Core principles and mechanisms: what kind, why, and when variability matters

Four kinds of variability

In reviewing the literature, we were surprised to discover that the label ‘variability’ has been used to refer to at least four different types of variability, each stemming from a different source. This is crucial to recognize, as these different ‘variability types’ may or may not have the same impact on learning and/or generalization. Across experiments and fields, low and high variability appear to be contrasted in four different ways (Figure 2): (i) numerosity (set size), such as when learning from more or fewer distinct examples; (ii) heterogeneity (differences between examples), such as when learning from a fixed number of examples that are more or less similar to one another (this similarity, in turn, can be along task-relevant or -irrelevant dimensions, as we discuss later); (iii) situational (contextual) diversity, such as when learning from the same examples under more or less variable environmental conditions that do not pertain to the examples.

Box 1. Useful variability as a critical factor in the success of machine learning

Artificial neural networks, loosely inspired by parallel, distributed, hierarchical information processing in the brain, have long been used as cognitive models [168]. In the past 10 years, the increase in processing power, optimization of learning algorithms, and, perhaps most importantly, increase in the size of training sets, has increasingly produced super-human performance on tasks ranging from object classification [169], to face recognition [170], to language processing [171], to playing games such as chess and Go [172].

A perennial problem in training neural networks (indeed, a problem general to any algorithm tasked with learning to associate stimuli with a response) is how to appropriately generalize from the training data [173]. Learning that is overly specific to the training set yields progressively poorer performance on new items (so-called overfitting, Figure 1A), limiting the usefulness of the algorithm to correctly respond only to items that very closely resemble those that had been presented in training [140]. One way to avoid overfitting is by using data augmentation to artificially increase the variability in the training set. In the visual domain, this has been done by rotating training images, changing their size, color balance, and by partially masking the objects of interest, etc. (Figure B). This variability is ordinarily part of normal human experience (e.g., we regularly see objects under different lighting conditions, from different perspectives, partially hidden behind other objects, etc.). Incorporating such variability into the training experience enables more robust and human-like performance in image recognition and classification (e.g., [24,136–138,141], speech recognition (e.g., [139,142–144]), and musical feature extraction (e.g., [145,146]). By introducing variation along nondiscriminative dimensions, data augmentation enriches the available input and helps the model learn the discriminative invariances (i.e., learn which dimensions most reliably predict the category across a wide range of contexts), leading to broader generalization.

Deciding the dimensions along which to increase variability is done in a largely haphazard way. Leveraging the insights from cognitive science of what kinds of variability matter and when variability is most beneficial may help machine learning researchers to construct more effective training sets. But given the relative ease of training neural networks compared with training people, perhaps a more likely possibility is the use of neural network models (and machine learning more generally) to gain insights into what kinds of variability may be most useful for each domain, an instance of machine teaching [174] wherein machine learning algorithms are tasked with producing training regimes that maximize learning efficacy.
themselves; and (iv) scheduling (e.g., interleaving, spacing), such as when learning from the same examples, but under more or less varied practice schedules (e.g., that differ in the order in which examples are presented or the time lag between them).

For instance, a hypothetical study comparing the effect of training tennis serves repeatedly from just one location (e.g., 6 inches to the right of the center mark), versus training from four different locations (e.g., 6, 7, 8, and 9 inches to the right of the center mark) would be testing the impact of numerosity (note that while it also technically manipulates heterogeneity, these are not intractably confounded, see later). A study that contrasted learning from four locations that are quite close to one another (e.g., 6, 7, 8, and 9 inches to the right of the center mark) with training on four

![Figure I. Variability effects in machine learning. (A) A typical relationship between error on training items (unbroken lines) and generalization error (broken lines) under less variable training (black) or more variable training (red). Performance in neural networks, as in people, is generally superior on training items (unbroken lines), showing that there is a cost to generalization. Notice how the error for the generalization items gradually starts to increase when training is less variable (black broken line). Artificially increasing variability, such as through data automatic augmentation techniques shown in (B) can improve generalization performance, but at the cost of slowing down initial learning (compare the unbroken red and black lines).](image-url)
Trends in Cognitive Sciences

Box 2. When ‘learning’ means ‘learning to generalize’

One major difficulty in extracting global principles across domains with regard to the tradeoff between variability and learning/generalization is that different fields and different scholars vary in how much they equate the term learning with generalization.

In some domains, the measure of interest is the degree to which the exact trained task is learned (e.g., in perceptual learning, the sole measure of interest might be the best performance that can be reached via repeated practice with one particular type of visual stimulus) and thus generalization performance in these domains might not be examined at all (see also Figure 3 in main text). Of the studies cited in this paper, approximately 30% do not examine generalization and focus only on the retention of knowledge gained during the training phase, without examining its transfer to other contexts. In other domains, however, the learning of the exact trained stimuli is uninteresting or even a nuisance to be controlled for in examining generalization (e.g., by training all participants to criterion levels of categorization accuracy on the training set and then testing generalization accuracy [15]).

These differences in focus often reflect the real-world importance of learning to generalize. In situations where there is little variation in the real world, there is simply no need to generalize beyond the trained data. This is the case in assembly lines at factories, for example, where people need only master a very specific skill and there is no need for them to be able to generalize this skill since the task is always the same. Likewise, when learning to read, children must be able to generalize letter-forms across fonts and sizes (see Figure 5 in main text) but do not need to generalize sound-to-form mappings beyond the specific writing system they are learning. Most real-world tasks, however, are not as invariant as factory assembly lines, and most domains (e.g., categorization, visual perception, motor learning) largely equate proper learning of a task with the ability to generalize. It is in these domains where the learning of specific trained stimuli is often of less interest. For example, if one’s command of spoken English was so specific that it is limited to understanding only a single speaker, we would hesitate to say that this person had full knowledge of the language. Likewise, we would not consider a child to have learned the category dog if they are unable to categorize a new dog as a dog, nor would we say that someone has really learned how to drive if they are unable to drive at different speeds, in a different city, or in a different car. More uniform approaches to reporting both training and generalization outcomes would help to uncover cases where the tradeoffs between learning the training set and generalizing are absent, or especially strong.

locations that are more spread apart (e.g., 6, 12, 18, and 24 inches to the right of the center mark) would be testing the effect of heterogeneity, while keeping numerosity constant. A study that contrasted learning to serve on a court painted green versus learning to serve on courts painted a variety of colors (e.g., red, blue, etc.) would be testing the impact of situational/contextual diversity. Finally, a study that contrasted different practice schedules (e.g., 6, 6, 6, 12, 12, 12, and 12 inches versus 6, 12, 6”, 12, 6, 12, 6, and 12 inches) would be examining the impact of order.

No single study to our knowledge has directly contrasted the effects of these four different sources of variability on the same target behavior, as in the earlier example. Instead, different studies typically attempt to tackle a single source of variability. For instance, many studies have specifically focused on the consequences of different training schedules: comparing ‘blocked’ training (e.g., AAA, BBB) with ‘interleaved’ training (e.g., BAC, ACB), or comparing ‘massed’ training (e.g., when learning events occur in succession) with ‘spaced’ training (e.g. when learning events are distributed over time). A typical finding is that interleaved and spaced training (which are considered to be more variable) lead to better learning and broader transfer of motor skills (e.g., volleyball serves [27,28], badminton serves [29], golf strokes [30], pistol shooting [31]) and of novel categories (e.g., trivia facts [32], toys [33], scientific concepts [34]). Other studies have focused specifically on the third source of variability, that is, situational variability, showing that variation in the external learning conditions such as the physical environment in which learning takes place affects performance. For example, memory for object labels is better when objects are displayed on different colored backgrounds as opposed to only a white background [35,36], and taking a class in different rooms leads to superior retention of class material compared with learning in the same room [37,38].

Differentiating between the first (numerosity) and second (heterogeneity) sources of variability is less common. Although some studies have focused specifically on heterogeneity by contrasting more or less diverse sets of exemplars while keeping set size constant (e.g., [39–42]), many other
studies have treated numerosity and heterogeneity as more-or-less interchangeable, effectively confounding them (but see [21] for a theoretical account that does explicitly differentiate them and [43–45] for unique attempts to experimentally tease them apart). As no two experiences are identical, having more examples to learn from indeed typically implies experiencing greater variability. For instance, experiments investigating effects of variability on speech perception often contrast exposure to one versus multiple speakers, with the number of speakers taken as a proxy for the amount of phonetic variability in the input. Yet, this experimental manipulation is not really about the number of speakers, but rather about the assumption that different speakers would produce more variable pronunciations compared with a single speaker (as demonstrated in [46]). Critically, numerosity and heterogeneity are not intractably confounded: in principle, a small set of exemplars can be much more heterogeneous than a large but rather similar set of exemplars and vice versa. Returning to the tennis example, practicing serving from fewer yet more far apart locations on the court would likely lead to broader generalization performance compared with when practicing from multiple but very close locations on the court, despite the latter including more training locations. Likewise, when learning to identify a novel category of

Figure 2. A schematic of the four different sources of variability. In the case of having one versus multiple examples during training, variability stems from numerosity. (Note that the degree of difference in color represents the degree of difference between examples.) In the case of training on a more or less diverse set of examples, variability would stem from heterogeneity. In many cases in the literature to date, numerosity and heterogeneity have been confounded, comparing the impact of training with many unique examples with training on a single example. However, as we see in the figure, these can be at least partially separated (see main text).
animals, exposing children to multiple similar exemplars is less effective than exposing them to a small yet diverse set of exemplars [47]. Thus, although numerosity is often taken as a proxy for heterogeneity, these two sources of variability do not necessarily have to align and it is often not the number of items or experiences per se that drive variability benefits. The only three studies that directly tried to disentangle the effects of numerosity and heterogeneity on categorization [43] and grammar learning [44,45] suggested that the main predictor of learning and generalization was the heterogeneity of the training examples and their statistical coverage of the to-be-learned behavior, rather than just the number of examples. In other words, while numerosity is often treated as a source of variability, it is likely that the more relevant source of variability is the diversity of the examples (i.e., heterogeneity), with which it is often confounded.

Yet, while the aforementioned examples are excellent guides for how one might identify the specific source of variability (e.g., numerosity vs. heterogeneity) that impacts learning, it is important to note several issues that make the consideration of variability a challenge across fields. Indeed, the simple questions of what counts as different/unique examples or contexts and how to best quantify the degree of difference across stimuli is not always straightforward. Metrics may, for example, vary substantially based on how learners (and researchers) understand the task and the dimensions that the task may meaningfully vary across in the real world. After all, the amount of variability learners typically encounter varies drastically across domains (i.e., the range in which examples can vary is different to begin with), as does the level of abstraction that maximizes real-world performance (Box 2). Take the dimension of size. In the real world, tennis courts do not vary in size, while the size of soccer pitches does vary. This difference across the sports will certainly impact the extent to which the size dimension is considered to be a dimension that might have some importance for generalization (and thus for training, noting this belief might not be accurate; there could be value in training even along seemingly totally ‘irrelevant’ dimensions, see the Mr Miyagi principle, Box 3). Critically though, the difference across sports could also impact the extent to which conditions that vary along this dimension are perceived as being different from one another. If researchers created three different tennis courts and three different soccer pitches, with the steps between each of the versions matched in terms of raw perceptibility (e.g., in just-noticeable difference steps), it may nonetheless be the case that individuals perceive training on the three differently sized tennis courts as being more variable than training on the three differently sized soccer pitches. This presents a challenge to researchers, as it is unclear whether the ‘raw’ or ‘perceived’ variability is more important to utilize as a measure.

Yet, in the absence of a uniform metric for quantifying differences between stimuli, it is difficult to say whether two conditions across domains (e.g., language vs. motor learning) or even within a domain (e.g., different words vs. different phonemes) are indeed matched in terms of heterogeneity. As such, work on the effect of variability might only be able to match stimuli in terms of the directionality, ratio, or type of variation, but not in terms of absolute magnitude.

Why does variability impact learning?

Across the different sources of variability and domains discussed earlier, the main phenomenon is largely the same: more variability initially hinders learning, but in many cases subsequently benefits generalization/transfer. This phenomenon has been articulated under multiple theoretical frameworks over the years, resulting in a range of competing and/or complementary theories across domains (Figure 3). While some of these theories share striking commonalities, they also differ substantially on multiple aspects. Specifically, different theories vary in their primary focus (i.e., explaining variability effects for learning trained stimuli, for generalization to novel stimuli, or for both), in the type of variability they consider or attempt to explain (e.g., some domains have largely focused on heterogeneity of inputs, while others are concerned only with variability in the temporal order with which inputs are presented), in the type of information that they assume
learners store in memory (i.e., whether it is specific events, abstractions over encountered events, or both), and, most importantly, in the underlying mechanisms at play (e.g., whether the mechanism underpinning the impact of variability is primarily contrastive in nature, whether it concentrates on the coverage of the to-be-learned space, etc.).

Notably, no two theories are identical, not even when they evoke the same mechanism for explaining the effects of variability. Illustrating this point, let us compare different theories on practice schedule variability (and specifically, spacing) in three different domains: categorization, list memorization, and motor learning. Across domains, the effects of variable schedules have been explained in terms of forgetting and reconstruction [33,48–50], with the main argument being largely the same: spaced training results in more forgetting, which forces learners to perform some form of active reconstruction when encountering the next event. However, the three theories differ in their focus and in their underlying assumption of what information is being coded: the forgetting theory in motor learning focuses solely on explaining variability effects in generalization and assumes that learners only store abstractions of motor functions, namely schemas, with specific events being stored only temporarily. The forgetting theory in verbal list memorization, however, focuses solely on explaining variability effects on learning the trained stimuli and assumes that learners store the specific events they encounter and only them. The forgetting theory in categorization incorporates the two, but adds an additional dimension in which variability in presentation order not only benefits retrieval, but also promotes more abstraction over events.

Box 3. The Mr Miyagi principle: when variability along ‘seemingly unconnected’ dimensions is helpful

In the 1984 classic movie Karate Kid, Mr Miyagi begins young Daniel’s karate instruction by having Daniel wash and wax cars: ‘Wax on, wax off’. Daniel is understandably frustrated; ‘Four days I’ve been busting my ass, I haven’t learned a thing’. Mr Miyagi disagrees: ‘You learned plenty… Not everything is as it seems.’ The moral, of course, is that the training, which on the surface seems completely disconnected from martial arts, has nonetheless been preparing Daniel in ways he does not appreciate (https://tvropes.org/pmwiki/pmwiki.php/Movie/WaxOnWaxOff). This idea has clear touchpoints with ideas in machine learning (e.g., in the form of the bias-variance tradeoff, wherein variability optimizes the output for inputs beyond those that have been observed and, in doing so, may produce a final outcome that is more robust to things like sampling error). Interestingly, this idea also has a long history in educational institutions. For example, the compulsory learning of Latin in European schools was often supported by arguments that learning its grammar promotes logical thinking [175]. More recently, analogous arguments have been made for continuing to teach children cursive handwriting. Although few use it into adulthood, some have argued that cursive promotes general fine-motor skills (https://www.nytimes.com/2014/06/03/science/whats-lost-as-handwriting-fades.html; https://www.nytimes.com/2011/04/28/us/28cursive.html).

What all these claims have in common is the idea that sometimes it may be better to practice not the skill itself, but its core component or an adjacent skill, with the effect of improving transfer to what a learner is actually interested in, what we call the Mr Miyagi principle.

The validity of specific claims can only be settled through empirical tests. Our bet is that for the examples above, students of martial arts, logic, and fine-motor skills would be better served by learning the actual skill they are interested in learning. At the same time, it would be wrong to reject the Mr Miyagi principle altogether. What is needed is a way of predicting when practicing a seemingly unconnected skill will produce better transfer than practicing the specific task. Doing this successfully requires a theory of which dimensions of variation, however irrelevant-seeming, are in actually relevant. For example, if the lighting conditions under which we need to catch a ball vary, it makes sense to practice catching at different times of day. The idea that practicing catching a ball or shooting a hockey puck illuminated only by flashing strobes would improve performance is odd given that we never have to generalize to those types of conditions. And yet, some evidence suggests that such practice is indeed more helpful than practice in regular conditions because it forces the learner to be more predictive than reactive in their movements [176,177]. Our folk intuitions about order of practice are also often mistaken. People often incorrectly assume that blocked/massed practice leads to better learning [87] and when given the choice, tend to choose massed rather than interleaved practice, to their detriment [178]. Finally, it is possible that the virtue of different types of training might depend heavily on the amount of training one receives. If an individual only has the opportunity to receive 30 minutes of training, it might be more effective to spend that time practicing actual martial arts blocks. If one has the opportunity to receive a great deal of practice, there might be virtue in adding these ‘seemingly unconnected’ bouts of practice.
Examining the different theories presented in Figure 3, it is useful to distinguish between three nonmutually exclusive reasons for why variability might impact learning and generalization: highlighting relevant task-dimensions, providing broader coverage, and boosting retrieval.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Theory name [* selected references]</th>
<th>Focuses on</th>
<th>Variability type</th>
<th>Stored in memory</th>
<th>What does variability affect?</th>
<th>Why does variability increase the breadth of generalization?</th>
</tr>
</thead>
<tbody>
<tr>
<td>categorization</td>
<td>Similarity/Prototype models [15,151,152]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights similar features</td>
</tr>
<tr>
<td>categorization</td>
<td>Exemplar theory [153–157]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage, highlights similar features</td>
</tr>
<tr>
<td>categorization</td>
<td>Rule-based/Feature characteristics models [93,158,159]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights similar features + taking their role/weight into account</td>
</tr>
<tr>
<td>categorization</td>
<td>Category density model [92]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage</td>
</tr>
<tr>
<td>categorization</td>
<td>Distribution in time [65]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage</td>
</tr>
<tr>
<td>categorization</td>
<td>Forgetting/Reconstruction theory [33]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Forces active retrieval + highlights similar features</td>
</tr>
<tr>
<td>categorization</td>
<td>Bayesian inference [21,160]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage + more confidence</td>
</tr>
<tr>
<td>categorization</td>
<td>Similarity coverage model [42]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage + more confidence</td>
</tr>
<tr>
<td>categorization</td>
<td>Attentional bias framework [55,56]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights similarities and differences due to comparisons</td>
</tr>
<tr>
<td>list memorization</td>
<td>Consolidation theory [69,161,162]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Provides time to consolidate</td>
</tr>
<tr>
<td>list memorization</td>
<td>Encoding variability theory [67,68]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage + more traces</td>
</tr>
<tr>
<td>list memorization</td>
<td>(In)Attention theory [163]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader attention</td>
</tr>
<tr>
<td>motor learning</td>
<td>Schema theory [103]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>More robust abstraction</td>
</tr>
<tr>
<td>motor learning</td>
<td>Forgetting/Reconstruction hypothesis [50]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Forces active retrieval</td>
</tr>
<tr>
<td>motor learning</td>
<td>Elaborative/Deeper processing hypothesis [54]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights similarities and differences due to comparisons</td>
</tr>
<tr>
<td>language</td>
<td>Semantic distinctiveness [127]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage + more traces</td>
</tr>
<tr>
<td>language</td>
<td>Bayesian inference [47]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Broader coverage + more confidence</td>
</tr>
<tr>
<td>language</td>
<td>Exemplar theory [59,60,164]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights similar features + more activation</td>
</tr>
<tr>
<td>language</td>
<td>Associative learning [52]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights and strengthens relevant/contrastive features while disregarding irrelevant/non-contrastive features</td>
</tr>
<tr>
<td>Language / Inductive reasoning</td>
<td>Connectionist models [165–167]</td>
<td>X X X X X X X</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td>Highlights and strengthens relevant/contrastive features while disregarding irrelevant/non-contrastive features</td>
</tr>
</tbody>
</table>

Figure 3. A comparison of variability theories in different research fields. Because not all theories perfectly match the divisions in the table below, dark gray boxes mean a strong match, whereas light gray boxes denote a somewhat weaker match. See [151–167].
Variability helps in identifying task-relevant dimensions and establishing correct decision boundaries. When posed with categorizing novel stimuli, we must learn which differences are relevant to category memberships and which are not [3,13,14,25,46,51–53]. For example, color is useful for distinguishing between lemons and limes, but not for distinguishing cars from trucks. Greater variation can help learners identify task-relevant and -irrelevant features and code their acceptable boundaries. For instance, practicing a tennis serve under variable conditions would highlight the common principles of this physical action (e.g., the muscles used), while underscoring that force, speed, and position can vary within a range. Similarly, infants exposed to different specimens of a novel animal category may learn that this species has a common shape, but that its size or color can vary within a specific range [3]. Such inferences will be further strengthened if learning situations have hierarchical structure that can be exploited (e.g., if an athlete has practiced other sports with similar swinging actions as a tennis serve or if an infant has encountered other novel animal categories where category members have a common shape, but differ in color and size).

One way in which low variability during learning can hinder the breadth of generalization is that exposing a learner to too few instances (or many instances that do not vary in the right ways) increases the probability that the experienced items are not representative of the category and thus are not adequate for identifying which properties predict category membership. We illustrate the problem in Figure 4 using a simple toy example of a 2D category-learning task in which a learner is presented with exemplars from two categories and has to learn a decision boundary that separates them. While presenting the learner with all exemplars (Figure 4A) should yield an optimal decision boundary, it comes at the cost of slower initial learning and is often unrealistic (learners typically do not have access to all exemplars at the time of learning). Training that includes low variability along the category-diagnostic dimension (Figure 4B) may lead to learning the incorrect decision boundary (i.e., a failure to identify what dimension is most diagnostic of category membership, such as learning that color is not important for distinguishing lemons and limes because one only observed unripe lemons). Increasing the number of presented stimuli while keeping variability along the category-diagnostic dimension low (Figure 4C) only increases confidence in the incorrect solution. However, the same small number of training items that are sufficiently variable along the category-diagnostic dimension can lead to a more appropriate decision boundary (Figure 4D).

Multiple theories in the fields of categorization, language, and motor learning evoke this type of mechanism and argue that the benefits of greater variability on generalization lie in identifying category-relevant features, dimensions, or relations. These theories typically pertain to variability stemming from increased numerosity and/or heterogeneity of the training set. For example, associative learning models of language learning suggest that exposure to more variable word pronunciations ensures that noncontrastive cues such as pitch do not become strongly linked to word identity [52]. High variation in acoustic feature X (e.g., pitch) across different pronunciations can signal to learners that feature X is not directly relevant to identifying the word, and can therefore be potentially ignored or abstracted over. At the same time, high variation in acoustic feature X can highlight the existence of other acoustic features (e.g., voicing) that, in contrast to feature X, exhibit little variance across different pronunciations, essentially signaling to learners that these other features may be crucial. Interestingly, some theories evoke similar reasoning for explaining training schedule effects. For instance, highlighting the comparative affordance of variability has also been used to explain the benefits of interleaved versus blocked training in motor learning (the elaborative processing hypothesis [54] and in categorization (the attentional bias framework [55,56]), although whether blocked or interleaved learning is more effective also depends on the structure of the categories [56]. These theories suggest that variable practice schedule can emphasize potential distinctions between similar variations of the same basic action or category,
which in turn leads to a more comprehensive representation and the embellishment of task-relevant information. Specifically, more variable presentations may encourage discrimination and differentiation between training events, while blocked training may encourage learners to identify the similarities between them.

**Variability gives greater coverage of task-relevant space: from extrapolation to interpolation**

People have a surprising tendency to generalize conservatively (i.e., to ‘hug the data’). Performance is generally always better on items seen during training compared with similar, but unseen items (e.g., [57]). This observation has been a major motivation for exemplar theories of concept learning, according to which the similarity between newly encountered and previously encountered items/events is of paramount importance [52, 58–60]. Crucially, generalization is often strongly predicted by typicality: we tend to generalize much better to unseen items that are typical compared with unseen items that are atypical [15], a form of an interpolation bias. For example, people’s classification of integers and polygons declines as the exemplars depart from the more typical ones: everyone classifies 400 as an even number and an equilateral triangle as a triangle, but many people mistakenly think 798 is an odd number and a sizable minority claim that
scalene triangles are not real triangles [61]. This difficulty in interpolating to atypical examples is considered one of the banes of education, that is, when students can successfully solve practice problems but are unable to generalize their solutions to new problems that are designed to measure the same principle [62–65]. Increasing variability during learning is one way to mitigate against overly conservative generalization by expanding the hypothesis space.

For instance, despite learning the correct decision boundary in Figure 4D,E, the absence of variability along the irrelevant/nondiagnostic dimension (Dimension 1) means learners would have to extrapolate their knowledge to exemplars at the extremes of Dimension 1. This extrapolation carries a cost: learners are likely to make errors or perform more slowly when extrapolating to unseen items (i.e., predicting the value of a data point lying outside of the observed range of data) compared with when interpolating to unseen items (i.e., predicting the value of a data point lying within the range of the observed data). Including training items that span the nondiagnostic dimension (Figure 4F) can mitigate this cost by essentially turning extrapolation into interpolation. The breadth of generalization (shown by the length of the line) and the certainty of the generalization (shown by the width of the gray error band) may be impacted differently by different learning experiences.

Multiple theoretical frameworks stress the idea that greater variability boosts generalization due to greater coverage. These theories span different sources of variability, namely numerosity, heterogeneity, contextual variability, and presentation schedules. For instance, Estes’s distribution in time theory on the effect of spacing suggests that breaking the temporal dependencies between events by increasing the variability of presentation order inherently produces more heterogeneity and more contextual/situational variability, seeing as events that are further apart in time are likely to be more different than one another [66]. This idea resonates with the encoding variability theory, which suggests that more time between training blocks creates a greater opportunity for general, contextual, and descriptive cues to change, increasing the likelihood that an item at test would be similar to one of the items seen during training [67,68]. More recently, Bayesian inference models of categorization and word learning suggest that learners update their beliefs about the likelihood of different probability distributions following exposure to specific examples. As a result, high heterogeneity in the set of observed examples leads to higher probability to generalize outside the examples’ range [21,47]. This principle also resonates with the idea that variability helps to approximate the real distribution in the world (i.e., if something is variable during learning, it is probably variable in similar ways outside of the specific learning task, see also Box 2).

**Better retrieval from memory**

Theories that focus solely on the effects of variable training schedules (i.e., order variability) also suggest that variability is linked to retrieval performance. In some theories, temporally variable training is argued to boost retrieval performance through a cycle of forgetting and reconstruction: if the same stimuli or motor action is repeated, the previous representation of it is still accessible in short-term memory and so there is no need to reconstruct it. However, if the same stimuli or motor action is repeated when the previous representation begins to fade or has already faded, it is necessary to go through a more effortful reconstruction process, rendering it more accessible next time [33,48–50]. In other words, while active reconstruction requires more effort (as it is easier to retrieve a previous representation of a stimulus that is still accessible in memory than to construct a new representation of a stimulus that has already faded from memory), this additional computational step improves retrieval performance.

Other theories explain the benefits of increased variability for retrieval in terms of consolidation: if the difference between the first and second repetition of the stimuli or motor action is longer, there
will be more time to consolidate it into long-term memory, which in turn strengthens retrieval [69]. Finally, the encoding variability theory suggests that learners benefit from spaced presentation since it increases the number and/or richness of memory traces and associative cues that can be used for retrieval and recall later on. In turn, associations with more variable cues increases the likelihood that, at test, the relevant cue will already be available, making retrieval easier [67,68,70].

**When is variability most helpful? Variability at different stages of learning**

Across domains, the exact effects of variability may depend on the stage of learning (i.e., whether the target behavior is familiar or new) and the type of variation people are exposed to (i.e., whether it is along discriminative or nondiscriminative dimensions) [25,51,71–73]. Specifically, variability can be more or less beneficial at different stages of learning.

In general, high variability can make learning more difficult when learners are in the very early stages of acquiring a target behavior. For instance, beginners and children who are only just ‘getting the hang’ of a motor skill (e.g., a tennis serve) or who are just being familiarized with a novel category benefit from low variability during initial practice (e.g., blocked training as opposed to more variable interleaved training; exposure to exemplars with little to no variation between them), and may experience difficulties or even get overwhelmed when too much variability is introduced at first [55,71,74–79]. For example, students with less prior knowledge who are learning to solve math problems benefit from receiving less variable examples first, while the opposite is true for students with more prior knowledge [63,80]. A similar benefit of strategically restricting early experience is found in the domain of progressive alignment and analogy, where starting with more concrete and less variable examples aids learning [81]. These findings are in line with the idea that learning may benefit from ‘starting small’ (i.e., that having less data or less complex data early on provides a learning advantage, e.g., better retention) [82,83]. This can also help reconcile previous conflicting findings that report differences in the effect of variability when testing children versus adults and between people with more versus less expertise.

The effect of high versus low variability in the early stages of learning also depends on the type of variability that learners experience, such as whether it is along discriminative or nondiscriminative dimensions (Figure 4). Variability along discriminative dimensions (i.e., dimensions that are useful for distinguishing the categories being learned or those that are causally linked to a target behavior) seems to impair learning by novices, while variability along nondiscriminative dimensions (i.e., nondiagnostic dimensions that are not linked to the target behavior) can actually promote learning, even in novice learners. For example, infants fail to discriminate between words that differed in the voicing of one sound (e.g., p or b) if they are exposed to variation along aspects of pronunciations that are directly relevant for differentiating between voiced and unvoiced consonants (e.g., voice onset time) [13]. That is, when infants are still in the process of establishing categorical distinctions based on voice onset time, variation along this discriminative feature hinders their learning of these categories because it makes it harder to detect how many categories there are and how these categories differ from one another [73]. But at the same time, infants can successfully differentiate between such minimal pairs when they are first exposed to variability along nondiagnostic aspects of the words’ pronunciation, such prosody, frequency, and vowel quality, which actually do not help to differentiate between these voiced and unvoiced consonants. That is, variability of nondiscriminative dimensions can help infants form robust and generalizable representations that include only phonetically relevant cues, while excluding irrelevant ones.

Whereas Figure 4 emphasizes differences in the kinds of variability the learner experiences, Figure 5 emphasizes when variability is experienced. If early training is insufficiently variable (Figure 5A), the hypothesis posited by the learner may become too narrow, making it difficult to
adjust it to subsequently encountered items that lie outside of the posited hypotheses. Encountering greater variability along nondiscriminative features early on (Figure 5B) helps ensure that the hypothesis about what makes something a category member remains sufficiently broad to include items that are likely to be experienced in the future. More variable training often also corresponds to more representative sampling of the true environmental variation. This in turn leads to broader coverage, increasing the likelihood that newly encountered examples/situations will be similar to previously encountered ones, allowing for interpolation rather than extrapolation. Greater variability need not entail more representative sampling, however. For example, the initial training set of the letter ‘A’ shown in Figure 5B includes very rarely encountered fonts, yet their inclusion may help the learner keep the hypothesis sufficiently broad to accommodate new exemplars that actually are encountered in the future. Surprisingly, few studies have contrasted introducing the exact same variability at different points in the learning process.

The ‘when’ component becomes particularly important in cases of active learning, where rather than passively waiting to be exposed to predetermined inputs, learners must seek out information for themselves. One way to generate additional information is by generating additional variability through exploration such as manual object manipulation [84] and pretend play [85,86] (see Outstanding questions). In such cases, early observations can impact how the learner subsequently explores the space: low initial degrees of variability could be taken as a signal that there is not much to discover and so there is little need to further explore. If variability is then introduced later on, it may be less effective at producing generalization.

Variability along discriminative features can also improve performance of more established behaviors later in learning. For example, adults who are exposed to non-native accented speech show better
adaptation to this accented speech when words vary along the relevant dimension of voice onset time [87]. In other words, the same type of variability that hindered infants’ learning of novel categories can be beneficial for proficient language users, who only need to tune their existing knowledge to successfully comprehend an unfamiliar non-native speaker. Similarly, more experienced mathematics students benefit from exposure to arithmetic problems with high rather than low variability [63]. Once essential problem-solving concepts and procedures have been acquired, learners are able to benefit from variable examples and perform better when exposed to a new problem. These findings show that in later stages when learners already have well-established categories or skills, variability along discriminative aspects can also facilitate learning.

Notably, even though the different effects of variability along discriminative versus nondiscriminative features have been identified separately in studies on language learning, motor learning, and categorization (e.g., [51,55,71,73]), many studies do not explicitly address this potential difference, despite its relevance for explaining why some variability manipulations are more successful than others (e.g., see Box 1 on data augmentation, which specifically varied nondiscriminative features without recognizing them as such).

Concluding remarks
The importance of variability for learning has been repeatedly rediscovered, renamed, and studied in multiple fields, with little acknowledgment of the overlap in findings and mechanisms (Table 1 and Figure 3). By placing different studies on variability alongside one another, it is possible to start seeing some of the core properties and general underlying principles.

Variability can arise from different sources: more training examples, more heterogeneous examples, more variable contexts, and more variable practice schedules. Strikingly, these four sources have not been explicitly defined and only rarely compared with one another, limiting our understanding of what kinds of variability are more effective and whether experience with more variability is fundamentally similar regardless of its source (e.g., with respect to its effects on strength versus breadth of generalization; see Outstanding questions). Moreover, even within those areas where a great deal of work has been conducted, it remains the case that large parts of the space remain relatively unexplored. For example, work on heterogeneity has tended to use distributions that are symmetric, such as uniform or normal distributions, rather than the types of skewed distributions that arguably better represent the way variability occurs in the real world in many domains [88]. Indeed, several studies have already looked at real-world distribution of variation by examining the effects of living in big versus small communities (Box 4).

Our review highlights several additional points. First, the effects of variability differ depending on both the learning stage and the features that are being varied; variability in discriminative versus nondiscriminative features and in late versus early stages of learning can yield different patterns of generalization. Second, comparing different theoretical accounts suggests that there are at least three nonmutually exclusive reasons for why variability might impact learning and generalization: variability helps to identify relevant task-dimensions, provides broader coverage, and boosts retrieval from memory. However, because different types of variability have rarely been directly contrasted, it remains unclear whether they involve shared or different mechanisms (see Outstanding questions). For example, it may be possible to capture the impact of both heterogeneity and scheduling via various inference-based processes [89]. Finally, the effects of variability may be fundamental enough to go beyond the brains or even the nervous system. For instance, more diverse microbial exposure in rural versus urban environments has been associated with a reduced risk for allergies and asthma in children [90] (see Outstanding questions), suggesting that
Box 4. Variability in social networks: living in large versus small communities

It is interesting to consider the effects of natural variability as they relate to the size and structure of a people’s social network. The available input in smaller and/or more tightly knit communities is often more restricted and homogeneous than the available input in larger and more sparsely connected communities [179–181] due to the typical confound between numerosity and heterogeneity discussed earlier. As a result, members of communities of varying sizes and degrees of connectivity may differ on behaviors that are sensitive to variability, including language and categorization.

A rapidly growing literature investigating the link between people’s social environment and learning suggests that this is indeed the case. For example, face recognition seems to be affected by whether people grew up in a small community (fewer than 1000 people) or in larger community (over 30,000 people): exposure to fewer faces during childhood was associated with diminished face memory and less specific neural response to faces compared with objects, suggesting that variability shapes not only behavioral abilities but also the functional architecture of the brain [182]. Another study found that people with larger social networks are better at perceiving vowels in a noisy environment [73], perhaps because of their exposure to more variable speech. People with more age-heterogeneous social networks have better lexical access (measured by how fast they are at naming pictures) and are more accurate in estimating how people of different ages would name objects [183]. Interestingly, there is also evidence that language complexity is affected by the size of the community, with people playing a communication game in larger groups developing more systematic and rule-based languages [184]. One explanation for this difference is that people in larger groups were exposed to more variable input, which promoted the formation of more generalizable grammars. There is also evidence that people’s tolerance to sexual nonconformity is affected by the size of the city they lived in during their teenage years: people who grew up in larger cities tend to be more tolerant to homosexuality, extramarital sex, premarital sex, and pornography [185]. Differences in experienced community structure may also be linked to differences in malleability of social stereotypes, which are also sensitive to perceived group variability: when a group is perceived as highly variable with respect to trait X, people are less likely to assign group membership to someone unfamiliar based on them having trait X [186]. Recent studies focusing on social networks formed by social media algorithms, which tend to pair users with like-minded users, show that doing so leads to echo chambers [187] with the power to further polarize people’s opinions [188].

variability may impact the entirety of our biological system, from the level of single cells to that of complex multicell systems such as the brain.

Acknowledgments
We wish to thank Jocelyn Parong and Bart de Boer for their helpful input. L.R. was partially supported by FWO project G0B4317N, S.G. was partially supported by Office of Naval Research grant N00014-17-1-2049, and G.L. was partially supported by NSF-BCS Award 2020969.

Declaration of interests
No interests are declared.

References
162. Wodelegren, W.A. (1972) Trace resistance and the decay of word and category memories. J. Math. Psychol. 9, 413–455