**Title:**

# How variability shapes learning and generalization

**Authors:** Limor Raviv[1,2,3], Gary Lupyan[4], Shawn C. Green[4]

**Affiliations:**

[1] LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[2] Centre for Social, Cognitive and Affective Neuroscience, University of Glasgow, Scotland

[3] Artificial Intelligence Lab, Department of Computer Science, Vrije Universiteit Brussels, Belgium

[4] Department of Psychology, University of Wisconsin-Madison, USA

**Email addresses:** limor.raviv@mail.huji.ac.il [corresponding author]

lupyan@wisc.edu

cshawn.green@wisc.edu

**ORCID:**

Limor Raviv: 0000-0002-0716-3553

Gary Lupyan: 0000-0001-8441-7433

Shawn C. Green: 0000-0002-9290-0262

## Abstract

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 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 vs. irrelevant dimensions, and examine the effects of introducing variability at different points in training.

### Trends/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.

**Keywords:**

Variability; Diversity; Learning; Generalization; Categorization; Language

## 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 literally 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].

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 (see 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 1 vs. 3 speakers [6], whether bean bag tosses are more accurate when practicing tossing from 1 vs. 3 locations [18], whether face recognition is more accurate when people are exposed to 1 vs. 4 different photos of an individual [19], and whether people become more

accurate in solving verbal statistical problems when trained on 1 vs. 4 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 (see 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 [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] (see 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 cross-talk 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) (see Box 2). Even in domains that focus on generalization, the term "generalization" has been used in several distinct ways: while some parts of the categorization literature primarily use generalization in the context of property induction (e.g., will learners generalize property X to a new exemplar?), other literatures use the term generalization more broadly to refer to the capacity to make effective decisions about new exemplars (e.g., which properties are diagnostic/relevant and which are not). Notably, here we adopt the latter, broader convention. In addition, there have been substantial differences across fields in the types of variability that have been emphasized during training, with little to no comparison of the effect of variability stemming from different sources. Moreover, learners' sensitivity to variability is often dependent on a number of factors that may vary across domains, such as the similarity between the learned stimuli and the transfer stimuli, learners' prior knowledge, etc.

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.

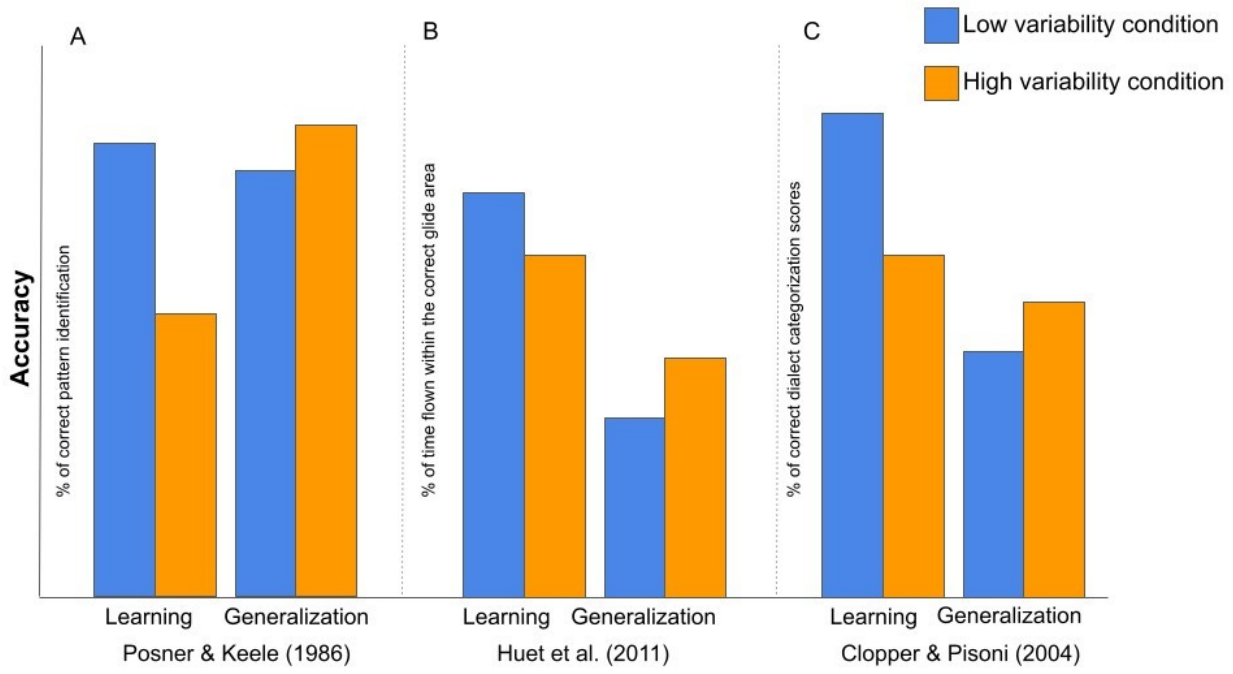


**Figure 1. Illustration of similar effects of variability across three domains**. A similar relationship between training stimuli and generalization has been observed in the domains of visual perception, motor learning, and language processing [4,15,17]. Bar plots were adapted from the reported results in the original studies. A: Posner & Keele (1968) measured the effect of variable training on the perception of visual patterns. Compared to 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. (2011) 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 & Pisoni (2004) 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 3 talkers in each dialect were better at recognizing the dialect of new, unfamiliar speakers (adapted from Figure 2).

**Table 1. The names given to the variability phenomenon in different research fields**

| Field | Term | Main Point | Representative Papers |
|---|---|---|---|
| Categorization / Visual perception | The Variability Effect | *"Variability improved classification of novel items, whereas repetition improved classification of studied items"* [91]<br><br>*"Infants can generalize a given property to new members of an animal category… only when presented with multiple exemplars of a familiarized category*" [3] | [3,15,16,19,39,40,74,91–93] |
| | Category Density | *"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]<br><br>*"A diverse category… led to lower levels of accuracy in the training task, wider generalization, and poorer item recognition"* [40] | [39-41,74-76,92-96] |
| | The Spacing Effect | *"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]<br><br>*"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] | [32-34,55,97-101] |
| Motor Learning | Variability of Practice | *"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] | [1,2,17,18,102-109] |

| | | | |
|---|---|---|---|
| | Contextual Interference Effect / Distribution of Practice Effect | *"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] | [27-31,71,110-114] |
| Language | Phonetic/Acoustic Variability | *"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] | [12,13,46,87,115-121] |
| | Talker Variability | *"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]<br><br>*"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] | [4-6,9,14,52,122-125] |

| | | | |
|---|---|---|---|
| | Contextual/Semantic Diversity | *"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] | [7,8,10,72,126-130] |
| | Intra-task Interference / Interleaving / Spacing Effect | *"[verbal] learning is powerfully affected by the temporal distribution of study time. Spaced (vs. massed) learning of items consistently shows benefits"* [101] | [101,131-135] |
| Computational Modelling / Deep Learning | Data Augmentation | *"Networks trained with heavier augmentation yield representations that are more similar between deep neural networks and the brain.. Larger variety during training may be more biologically plausible than training with constant images or very light transformations"* [136]<br><br>*"Data augmentation builds up the model's tolerance to noise so it can better generalize to new images in the test set"* [137] | [24,136-146] |
| | The Variability Effect | *"All other things being equal, the lower the variability in the set of observed examples, the lower the probability of generalization outside their range."* [21] | [21-23,58] |
| Inductive Reasoning | The Diversity Effect / Diversity Premise | *"The less similar CAT(P1)..CAT(Pn) [the categories mentioned in the premise sentences] are among themselves, the more P1...P2 [the premise sentences] confirm C [the conclusion sentence]."* [42] | [42,147,148] |

| | | | |
|---|---|---|---|
| Problem Solving | Variability of Worked Examples | *"Increased variability of practice... is beneficial to schema acquisition and hence to transfer of acquired skills… The confrontation with a wide range of different problems and solutions of these problems is important to give inductive processes the opportunity to extend or restrict the range of applicability. However, because practice-problem variability is positively related to cognitive load, … increased variability may also be expected to hinder learning"* [62]<br><br>*"A set of high variability tasks is intrinsically more difficult to complete compared to a similar set of low variability tasks… Learning and problem solving with high variability tasks are expected to improve because the quality of constructed knowledge is enhanced"* [63] | [20,62,63] |
| Education | Desirable Difficulties | *"When instruction occurs under conditions that are constrained and predictable, learning tends to become contextualized. Material is easily retrieved in that context, but the learning does not support later performance if tested at a delay, in a different context, or both. In contrast, varying conditions of practice - even varying the environmental setting in which study sessions take place - can enhance recall on a later test"* [149] | [149,150] |

## 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 across fields, low and high variability appear to be contrasted in four different ways (see Figure 2): (1) **Numerosity** (set size) such as when learning from more or fewer distinct examples; (2) **Heterogeneity** (differences between examples) such as when learning from examples that are more similar or less similar to one another (this similarity, in turn, can be along task-relevant or task-irrelevant dimensions, as we discuss below); (3) **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 themselves; and (4) **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” to the right of the center mark), versus training from four different locations (e.g., 6”, 7”, 8”, 9” 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 below). A study that contrasted learning from four locations that are quite close to one another (e.g., 6”, 7”, 8”, 9” to the right of the center mark) with training on four locations that are more spread apart (e.g., 6”, 12”, 18”, 24” 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”, 6”, 12”, 12”, 12”, 12” versus 6”, 12”, 6”, 12”, 6”, 12”, 6”, 12”) 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 example above. 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- 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 to learning in the same room [37,38].

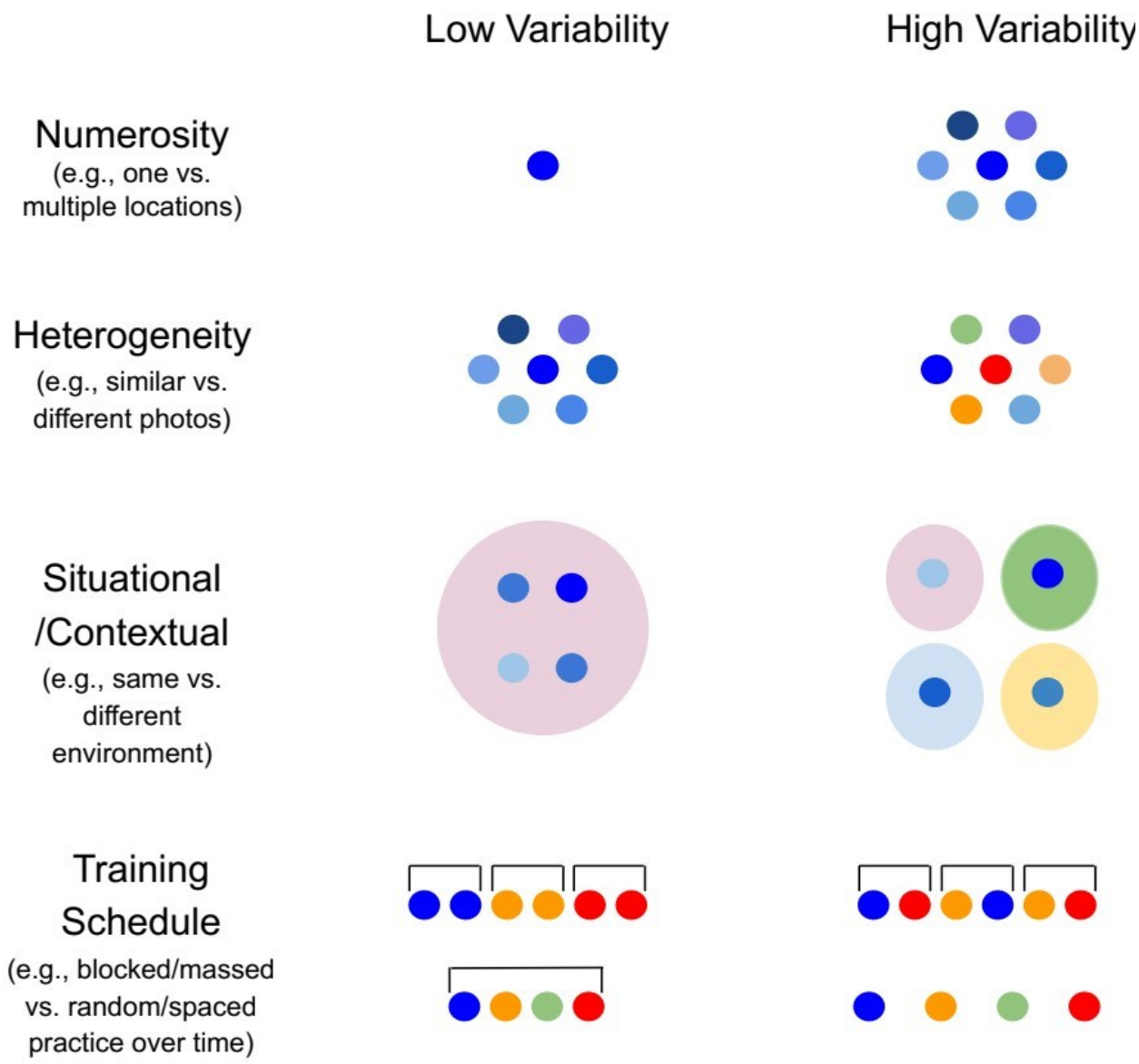


**Figure 2. A schematic of different sources of variability.** In the case of having one vs. multiple examples during training, variability stems from numerosity (*note that the degree of difference in color represents the degree of difference between examples). In the case of training on a similar vs. 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 (see main text). As we see though in the figure above, these can be at least partially separated.

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 typically implies experiencing greater variability. For instance, experiments investigating effects of variability on speech perception often contrast exposure to one vs. 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 to 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 to 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 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 two 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 examples above are excellent guides for how one might identify the specific source of variability (e.g., numerosity versus heterogeneity) that impact learning, it is important to note a number of 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 (see 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 above, 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 (see 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 in learning of trained stimuli, in generalization to novel stimuli, or 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.

Examining the different theories presented in Figure 3, it is useful to distinguish between three non-mutually exclusive reasons for *why* variability might impact learning and generalization: highlighting relevant task-dimensions, providing broader coverage, and boosting retrieval.

1. *Variability helps in identifying task-relevant dimensions and establish 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 task-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 so are not adequate for identifying which properties predict category membership. We illustrate the problem in Fig. 4 using a simple toy example of a 2-dimensional 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 (Fig. 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 (Fig. 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 isn't 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 (Fig. 4c) only increases confidence in the incorrect solution. On the other hand, the same small number of training items that are sufficiently variable along the category-diagnostic dimension can lead to a more appropriate decision boundary (Fig. 4d).

| Domain | Theory name [+ selected references] | Focuses on | | Variability type | | | | Stored in memory | | What does variability affect? | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Learning of inputs | Generalization | Numerosity | Heterogeneity | Contextual | Schedule | Abstraction | Specific items | Retrieval | Representation | Why does variability increase the breadth of generalization? |
| **categorization** | **Similarity/ Prototype models** [15,151,152] | | X | | X | | | X | X | | X | Highlights similar features |
| **categorization** | **Exemplar theory** [153–157] | | X | X | X | | | | X | | X | Broader coverage, highlights similar features |
| **categorization** | **Rule-base/ Feature characteristics models** [93,158,159] | | X | X | X | | | X | X | | X | Highlights similar features + taking their role/weight into account |
| **categorization** | **Category density model** [92] | X | X | | X | | | X | X | | X | Broader coverage |
| **categorization** | **Distribution in time** [66] | X | X | | X | X | X | X | X | | X | Broader coverage |
| **categorization** | **Forgetting/ Reconstruction theory** [33] | X | X | | | | X | X | X | X | X | Forces active retrieval + highlights similar features |
| **categorization** | **Bayesian inference** [21,160] | X | X | X | X | | | X | | | X | Broader coverage + more confidence |
| **categorization / inductive reasoning** | **Similarity coverage model** [42] | | X | | X | | | X | ? | | X | Broader coverage + more confidence |
| **categorization / inductive reasoning** | **Attentional bias framework** [55,56] | X | X | | | | X | X | X | | X | Highlights similarities and differences due to comparisons |
| **list memorization** | **Consolidation theory** [69,161,162] | X | | | | | X | | X | X | | Provides time to consolidate |
| **list memorization** | **Encoding variability theory** [67,68] | X | | | | X | X | X | X | X | X | Broader coverage + more traces |
| **list memorization** | **(In)Attention theory** [163] | X | | | | | X | | X | | X | Broader attention |
| **list memorization** | **Forgetting/ Reconstruction theory** [48,49] | X | | | | | X | | X | X | | Forces active retrieval |
| **motor learning** | **Schema theory** [103] | X | X | X | X | X | | X | X | | X | More robust abstraction |
| **motor learning** | **Forgetting/ Reconstruction hypothesis** [50] | | X | | | | X | X | | X | | Forces active retrieval |
| **motor learning** | **Elaborative/ Deeper processing hypothesis** [54] | X | X | | | | X | X | X | | X | Highlights similarities and differences due to comparisons |
| **language** | **Semantic distinctiveness** [127] | X | X | | | X | | X | | X | X | Broader coverage + more traces |
| **language** | **Bayesian inference** [47] | X | X | X | X | | | X | | | X | Broader coverage + more confidence |
| **language** | **Exemplar theory** [59,60,164] | X | X | | X | X | | | X | X | X | Highlights similar features + more activation |
| **language** | **Associative learning** [52] | X | X | | X | | | X | X | | X | Highlights and strengthens relevant/contrastive features while disregarding irrelevant/non-contrastive features |
| **Language / Inductive reasoning** | **Connectionist models** [165–167] | X | X | X | X | | | X | X | X | X | Highlights and strengthens relevant/contrastive features while disregarding irrelevant/non-contrastive features |

**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.

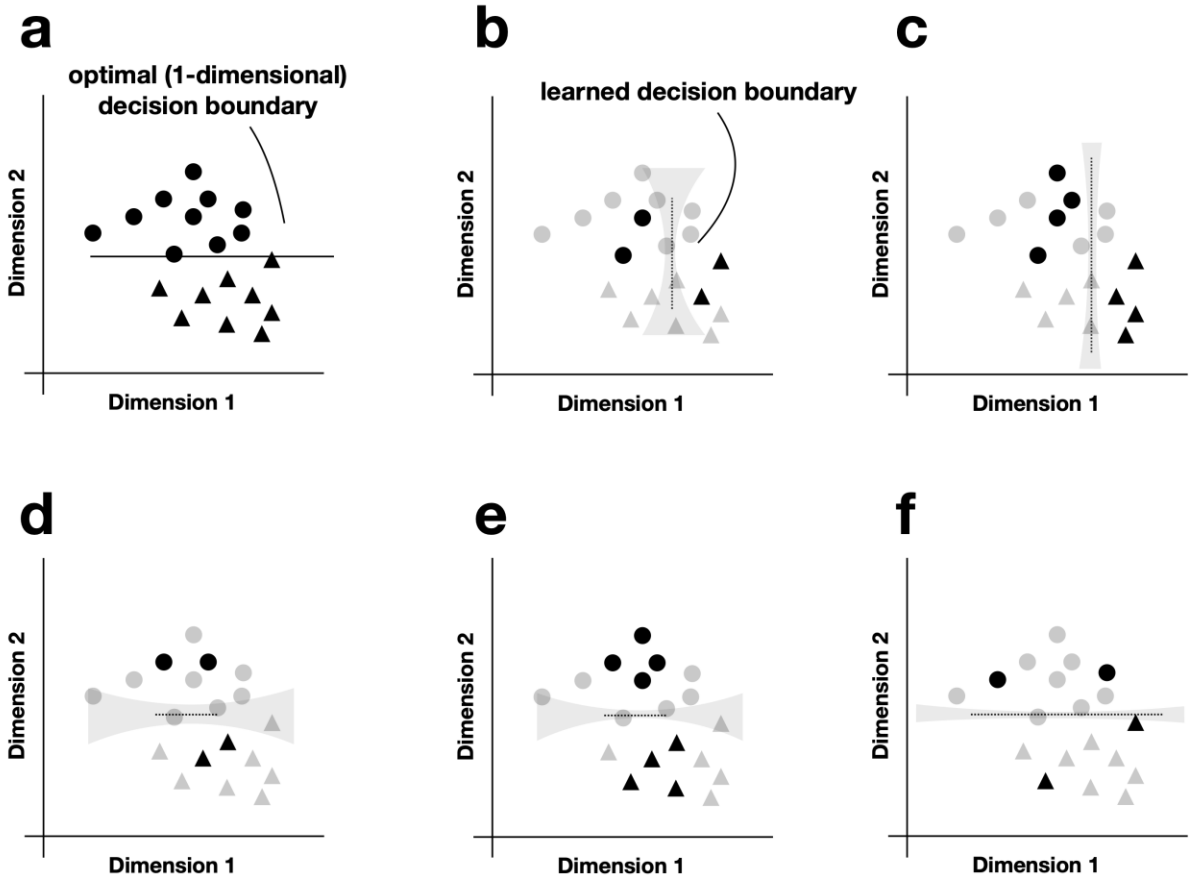


**Figure 4: Schematic of the effect of variability on learning a decision boundary.** An example of a simple categorization task in which Dimension 2 is relevant (where one is along this dimension determines category membership); Dimension 1 is not. Circles and triangles depict two to-be-learned categories. Black symbols show the examples available to the learner during initial training; gray symbols are unseen examples. When all possible examples are available (a), the ideal decision boundary is depicted by a solid black line. In all other panels, where only part of the possible set of examples are available, a black dotted line indicates high certainty in the boundary and the gray error ribbon represents the degree of uncertainty. Insufficient variability (b) and unrepresentative sampling (c) can lead to an entirely incorrect decision boundary. Greater variability along the diagnostic dimension (d) can lead to a correct decision boundary, but insufficient variability along the irrelevant dimension can lead to uncertainty in the slope of the decision boundary. This uncertainty remains if the learner simply observes more numerous examples (e), but shrinks when the variability spans both the diagnostic and non-diagnostic dimensions (f).

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 non-contrastive cues such as pitch do not become strongly linked to word identity [52]. High variation in acoustic feature X 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 (such as 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 vs. 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.

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

People have a surprising tendency to generalize conservatively – to "hug the data". Performance is generally always better on items seen during training compared to 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 typical items compared to unseen atypical items [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 even and an equilateral triangle as a triangle, but many people mistakenly think 798 is odd 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 Fig. 4d-4e, the absence of variability along the irrelevant/non-diagnostic dimension (Dimension 1) means learners would

have to extrapolate their knowledge (i.e., predict the value of a data point lying outside of the observed range of data) to exemplars at the extremes of Dimension 1; this extrapolation carries a cost, i.e., learners are likely to make errors or perform more slowly when extrapolating to unseen items compared to 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 non-diagnostic dimension (Fig. 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 (1955) 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 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).

*3. 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: 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 non-discriminative 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 *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); they may experience difficulties or even get overwhelmed when too much variability is introduced at first [55,71,74–79]. 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], and can help reconcile previous conflicting findings that report differences in the effect of variability when testing children vs. adults and between people with more vs. less expertise.

The effect of high vs. 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 non-discriminative dimensions (Fig. 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 *non-discriminative* dimensions (i.e., non-diagnostic 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* vs. *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 non-diagnostic 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 non-discriminative 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 (Fig. 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 early on (Fig 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 Fig. 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.

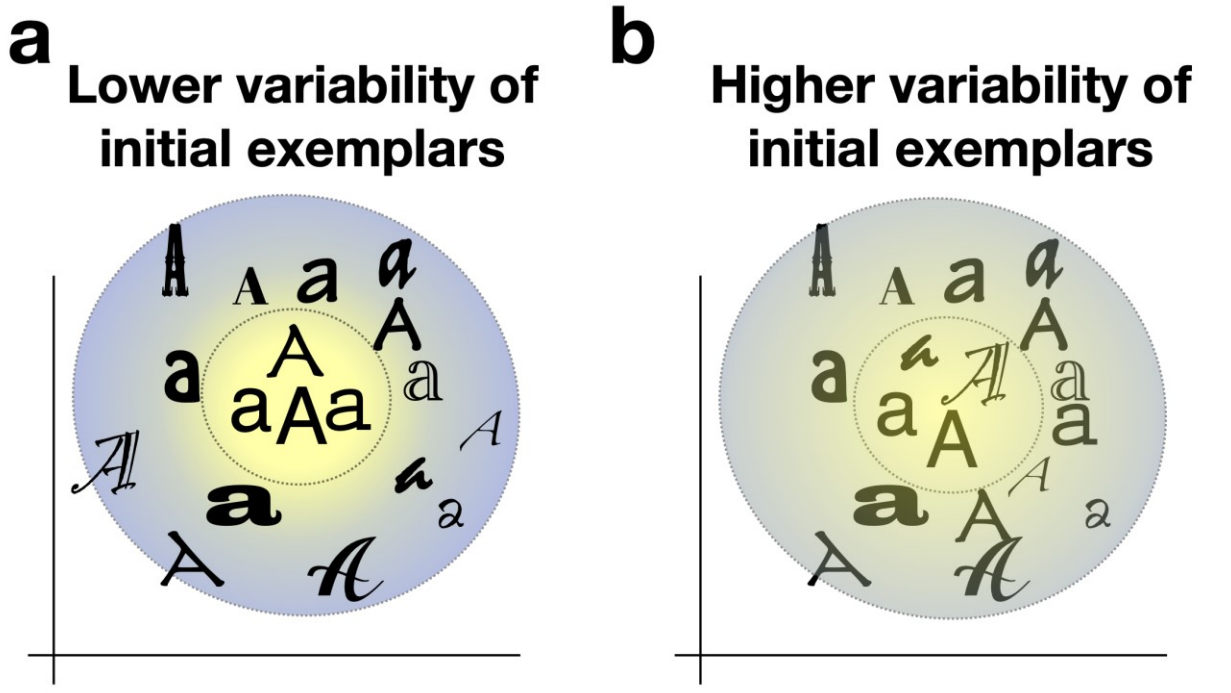


**Figure 5: Schematic of introducing variability during initial training and its impact on generalization.** An example of the effect of exposure to more or less variability when learners learn to identify what the letter "A" looks like. Initial training items are shown in the center of each panel. Color symbolizes

generalization performance: yellow is greater accuracy and/or certainty, and blue is lower accuracy and/or certainty. A less variable initial training set (panel A) can cause a learner to form a more specific hypothesis about what the letter "A" looks like, resulting in narrower generalization. To the extent that these early hypotheses can become entrenched, they can also limit future exploration. A more variable initial training set (panel B) can help keep the learner's hypothesis of what constitutes a letter "A" broader, allowing for more accurate and/or certain classification of exemplars encountered later.

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 is 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 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 was beneficial for proficient language users, who only needed 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 vs. non-discriminative 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 non-discriminative features without recognizing them as such).

## Concluding remarks

The importance of variability for learning has been repeatedly re-discovered, re-named, and studied in multiple fields, with little acknowledgment of the overlap in findings and mechanisms (see Table 1 and Figure 3). By placing different studies on variability alongside one another, it is possible to start to see some of its 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 to 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 vs. small communities (see 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 vs. non-discriminative features and in late vs. early stages of learning can yield complex patterns of generalization. Second, comparing different theoretical accounts suggests that there are at least three non-mutually exclusive reasons for why variability might impact learning and generalization: variability helps identifying 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, 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 vs. urban environments has been associated with a reduced risk for allergies and asthma in children [90] (see Outstanding Questions), suggesting that variability may impact the entirety of our biological system, from the level of single cells to that of complex multi-cell systems such as the brain.

**Outstanding questions:**

How does the brain handle variability? What exactly is being stored about individual experiences, and how much variation is encoded vs. discarded?

How domain-general is the variability effect? Are there different thresholds and/or transfer rates for different perceptual modalities and different tasks?

Do different sources of variability have different effects on behavior? For example, does contextual variability impact learning outcomes differently and/or lead to different degrees of generalization compared to heterogeneity or to varying training schedules?

Do the various types of variability differentially impact the strength of generalization (the certainty or consistency with which an individual generalizes learning to new items) versus the breadth of generalization (i.e., the maximum “distance” from the training set that the impact of learning is observed)?

Does the impact of variability differ in supervised learning (when externally or self-generated feedback is available) or unsupervised learning (when one is attempting to learn from the distributions alone)?

Is the relationship between variability and learning broadly similar across species or are there species-specific adaptations?

What are the mechanisms in which individuals introduce (i.e., self-generate) variability during learning? Specifically, what are the roles of pretend play, exploration (in contrast to exploitation), selective attention, and manual object manipulation?

How similar are the effects of variability in neural systems subserving cognitive/motor/perceptual functions to the effects of variability in other adaptive biological systems such as those underlying the immune response?

**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 last 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, Fig. IA), 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. (Fig. IB). 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 non-discriminative 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 may help machine learning construct more effective training sets. But given the relative ease of training neural networks compared to 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.

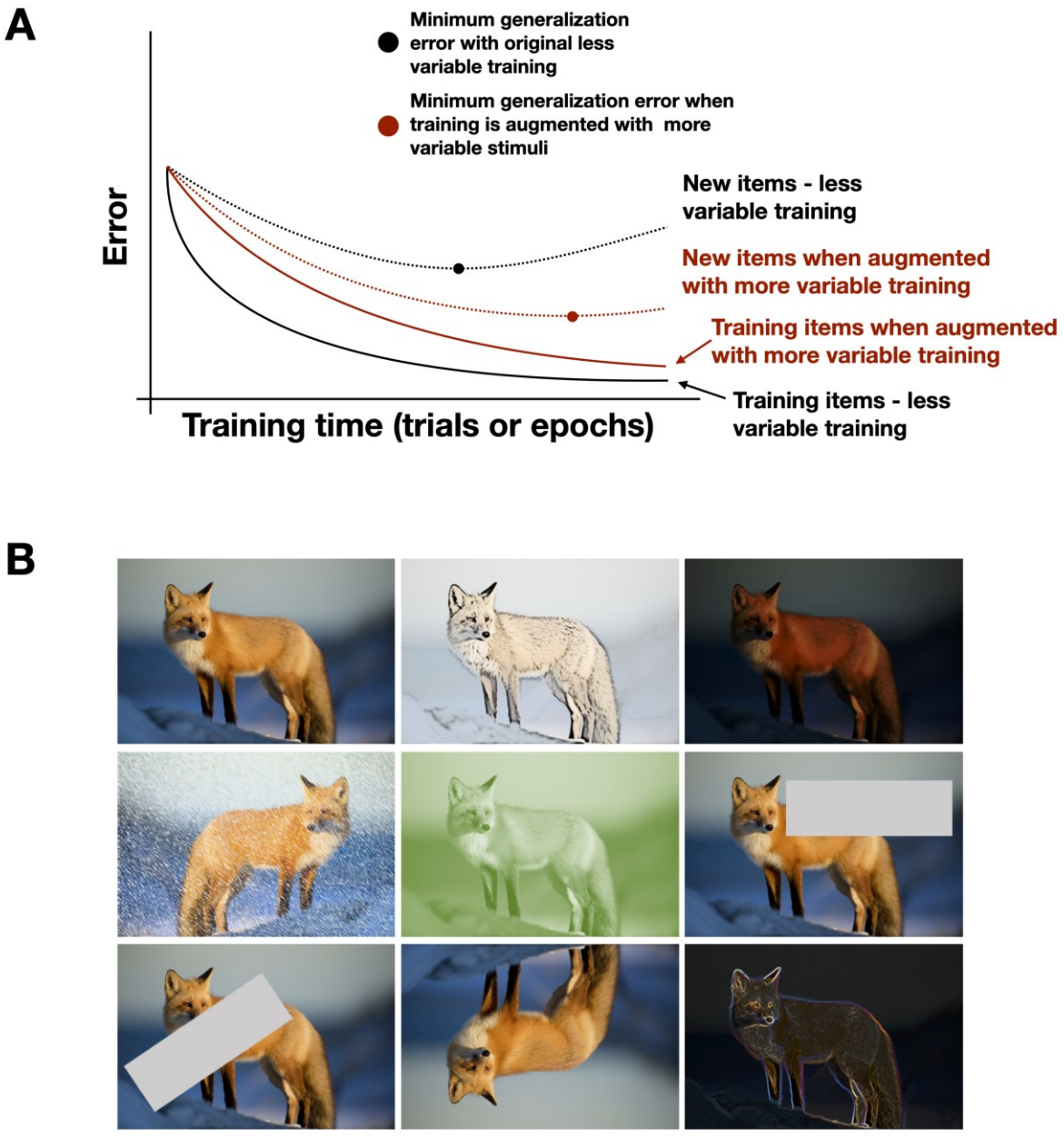

**Figure I (BOX 1). Variability effects in Machine Learning**. (A) A typical relationship between error on training items (unbroken lines) and generalization error (broken lines) under less variable training (black) or more variable training (red). Performance in neural networks, as in people, is generally 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).

### 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). Of the studies cited in this paper, approximately 30% have not examined generalization at all, focusing only on the retention of knowledge gained during the training phase, but not on 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., 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 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 as to be 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.

**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 learn[ed] 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 (see: https://tvtropes.org/pmwiki/pmwiki.php/Main/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 (see: 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 actuality 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 [97] and when given the choice, tend to mass 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.

**Box 4: Variability in social networks: Living in large vs. 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 input in larger and more sparsely connected communities [179–181] due to the 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 30k people): exposure to fewer faces during childhood was associated with diminished face memory and to less specific neural response to faces compared to 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 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].

**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.

## References


1 Douvis, S.J. (2005) Variable practice in learning the forehand drive in tennis. *Percept. Mot. Skills* 101, 531–545
2 Hernández-Davo, H. *et al.* (2014) Variable training: effects on velocity and accuracy in the tennis serve. *J. Sports Sci.* 32, 1383–1388
3 Vukatana, E. *et al.* (2015) One is Not Enough: Multiple Exemplars Facilitate Infants' Generalizations of Novel Properties. *Infancy* 20, 548–575
4 Clopper, C.G. and Pisoni, D.B. (2004) Effects of Talker Variability on Perceptual Learning of Dialects. *Lang. Speech* 47, 207–238
5 Seidl, A. *et al.* (2014) Talker Variation Aids Young Infants' Phonotactic Learning. *Lang. Learn. Dev.* 10, 297–307
6 Barcroft, J. and Sommers, M.S. (2005) Effects of acoustic variability on second language vocabulary learning. *Stud. Second Lang. Acquis.* 27, 387–414
7 Frances, C. *et al.* (2020) The effects of contextual diversity on incidental vocabulary learning in the native and a foreign language. *Sci. Rep.* 10, 13967
8 Johns, B.T. *et al.* (2016) The influence of contextual diversity on word learning. *Psychon. Bull. Rev.* 23, 1214–1220
9 Richtsmeier, P.T. *et al.* (2009) Statistical frequency in perception affects children's lexical production. *Cognition* 111, 372–377
10 Gómez, R.L. (2002) Variability and detection of invariant structure. *Psychol. Sci.* 13, 431–436
11 Eidsvåg, S.S. *et al.* (2015) Input Variability Facilitates Unguided Subcategory Learning in Adults. *J. Speech Lang. Hear. Res. JSLHR* 58, 826–839
12 Singh, L. (2008) Influences of High and Low Variability on Infant Word Recognition. *Cognition* 106, 833–870
13 Rost, G.C. and McMurray, B. (2010) Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy Off. J. Int. Soc. Infant Stud.* 15, 608–635
14 Rost, G.C. and McMurray, B. (2009) Speaker variability augments phonological processing in early word learning. *Dev. Sci.* 12, 339–349
15 Posner, M.I. and Keele, S.W. (1968) On the genesis of abstract ideas. *J. Exp. Psychol.* 77, 353–363
16 Hussain, Z. *et al.* (2012) Versatile perceptual learning of textures after variable exposures. *Vision Res.* 61, 89–94
17 Huet, M. *et al.* (2011) The education of attention as explanation of variability of practice effects: Learning the final approach phase in a flight simulator. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1841–1854
18 Kerr, R. and Booth, B. (1978) Specific and Varied Practice of Motor Skill. *Percept. Mot. Skills* 46, 395–401
19 Dukes, W.F. and Bevan, W. (1967) Stimulus variation and repetition in the acquisition of naming responses. *J. Exp. Psychol.* 74, 178–181
20 Quilici, J.L. and Mayer, R.E. (1996) Role of examples in how students learn to categorize statistics word problems. *J. Educ. Psychol.* 88, 144–161
21 Tenenbaum, J.B. and Griffiths, T.L. (2001) Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24, 629–640
22 Gliozzi, V. and Plunkett, K. (2018) Self-organizing maps and generalization: an algorithmic description of Numerosity and Variability Effects. *ArXiv180209442 Cs Q-Bio* at <http://arxiv.org/abs/1802.09442>
23 Hill, F. *et al.* (2019) Emergent Systematic Generalization in a Situated Agent. *ArXiv191000571 Cs* at <http://arxiv.org/abs/1910.00571>
24 Hernández-García, A. and König, P. (2019) Data augmentation instead of explicit regularization. *ArXiv180603852 Cs* at <http://arxiv.org/abs/1806.03852>
25 Boyce, B.A. *et al.* (2006) Implications for Variability of Practice from Pedagogy and Motor Learning Perspectives: Finding a Common Ground. *Quest* 58, 330–343

26 Schmidt, R.A. and Bjork, R.A. (1992) New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychol. Sci.* 3, 207–218
27 Bortoli, L. *et al.* (1992) Effects of contextual interference on learning technical sports skills. *Perceptualand Mot. Ski.* 75, 555–562
28 Travlos, A.K. (2010) Specificity and Variability of Practice, and Contextual Interference in Acquisition and Transfer of an Underhand Volleyball Serve. *Percept. Mot. Skills* 110, 298–312
29 Goode, S. and Magill, R.A. (1986) Contextual Interference Effects in Learning Three Badminton Serves. *Res. Q. Exerc. Sport* 57, 308–314
30 Porter, J. *et al.* (2007) The Effects of Three Levels of Contextual Interference on Performance Outcomes and Movement Patterns in Golf Skills. *Int. J. Sports Sci. Coach.* 2, 243–255
31 Keller, G.J. *et al.* (2006) Contextual Interference Effect on Acquisition and Retention of Pistol-Shooting Skills. *Percept. Mot. Skills* 103, 241–252
32 Cepeda, N.J. *et al.* (2008) Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychol. Sci.* 19, 1095–1102
33 Vlach, H.A. *et al.* (2008) The spacing effect in children's memory and category induction. *Cognition* 109, 163–167
34 Vlach, H.A. and Sandhofer, C.M. (2012) Distributing Learning Over Time: The Spacing Effect in Children's Acquisition and Generalization of Science Concepts. *Child Dev.* 83, 1137–1144
35 Twomey, K.E. *et al.* (2018) All the Right Noises: Background Variability Helps Early Word Learning. *Cogn. Sci.* 42, 413–438
36 Goldenberg, E.R. and Johnson, S.P. (2015) Category Generalization in a New Context: The Role of Visual Attention. *Infant Behav. Dev.* 38, 49–56
37 Smith, S.M. and Rothkopf, E.Z. (1984) Contextual enrichment and distribution of practice in the classroom. *Cogn. Instr.* 1, 341–358
38 Smith, S.M. *et al.* (1978) Environmental context and human memory. *Mem. Cognit.* 6, 342–353
39 Oakes, L.M. *et al.* (1997) By land or by sea: The role of perceptual similarity in infants' categorization of animals. *Dev. Psychol.* 33, 396–407
40 Hahn, U. *et al.* (2005) Effects of category diversity on learning, memory, and generalization. *Mem. Cognit.* 33, 289–302
41 Perry, L.K. *et al.* (2010) Learn Locally, Think Globally: Exemplar Variability Supports Higher-Order Generalization and Word Learning. *Psychol. Sci.* 21, 1894–1902
42 Osherson, D.N. *et al.* (1990) Category-based induction. *Psychol. Rev.* 97, 185–200
43 Bowman, C.R. and Zeithamova, D. (2020) Training set coherence and set size effects on concept generalization and recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 46, 1442–1464
44 Schiff, R. *et al.* (2021) Stimulus variation-based training enhances artificial grammar learning. *Acta Psychol. (Amst.)* 214, 103252
45 Poletiek, F.H. and van Schijndel, T.J.P. (2009) Stimulus set size and statistical coverage of the grammar in artificial grammar learning. *Psychon. Bull. Rev.* 16, 1058–1064
46 Galle, M.E. *et al.* (2015) The Role of Single Talker Acoustic Variation in Early Word Learning. *Lang. Learn. Dev.* 11, 66–79
47 Xu, F. and Tenenbaum, J.B. (2007) Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272
48 Challis, B.H. (1993) Spacing effects on cued-memory tests depend on level of processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 389
49 Jacoby, L.L. (1978) On interpreting the effects of repetition: Solving a problem versus remembering a solution. *J. Verbal Learn. Verbal Behav.* 17, 649–667
50 Lee, T.D. and Magill, R.A. (1985) Can Forgetting Facilitate Skill Acquisition? In *Advances in Psychology* 27 (Goodman, D. et al., eds), pp. 3–22, North-Holland
51 Ankowski, A.A. *et al.* (2013) Comparison Versus Contrast: Task Specifics Affect Category Acquisition: Comparison Versus Contrast. *Infant Child Dev.* 22, 1–23
52 Apfelbaum, K.S. and McMurray, B. (2011) Using Variability to Guide Dimensional Weighting: Associative Mechanisms in Early Word Learning. *Cogn. Sci.* 35, 1105–1138
53 Schyns, P.G. *et al.* (1998) The development of features in object concepts. *Behav. Brain Sci.* 21, 1–17; discussion 17-54

54 Shea, J.B. and Zimny, S.T. (1983) Context Effects in Memory and Learning Movement Information. In *Advances in Psychology* 12 (Magill, R. A., ed), pp. 345–366, North-Holland
55 Carvalho, P.F. and Goldstone, R.L. (2014) Effects of interleaved and blocked study on delayed test of category learning generalization. *Front. Psychol.* 5,
56 Carvalho, P.F. and Goldstone, R.L. (2015) The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychon. Bull. Rev.* 22, 281–288
57 Goldwater, M.B. *et al.* (2018) Relational discovery in category learning. *J. Exp. Psychol. Gen.* 147, 1–35
58 Rogers, T.T. and McClelland, J.L. (2004) *Semantic Cognition: A Parallel Distributed Processing Approach*, The MIT Press.
59 Johnson, K. (2006) Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *J. Phon.* 34, 485–499
60 Ambridge, B. (2019) Against stored abstractions: A radical exemplar model of language acquisition: *First Lang.* DOI: 10.1177/0142723719869731
61 Lupyan, G. (2013) The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition* 129, 615–636
62 Paas, F.G.W.C. and Van Merrienboer, J.J.G. (1994) Variability of Worked Examples and Transfer of Geometrical Problem-Solving Skills: A Cognitive-Load Approach. *J. Educ. Psychol.* 86, 122–133
63 Likourezos, V. *et al.* (2019) The Variability Effect: When Instructional Variability Is Advantageous. *Educ. Psychol. Rev.* 31, 479–497
64 Noble, T. *et al.* (2012) "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *J. Res. Sci. Teach.* 49, 778–803
65 McNeil, N.M. *et al.* (2006) Middle-School Students' Understanding of the Equal Sign: The Books They Read Can't Help. *Cogn. Instr.* 24, 367–385
66 Estes, W.K. (1955) Statistical theory of distributional phenomena in learning. *Psychol. Rev.* 62, 369–377
67 Melton, A.W. (1970) The situation with respect to the spacing of repetitions and memory. *J. Verbal Learn. Verbal Behav.* 9, 596–606
68 Glenberg, A.M. (1979) Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Mem. Cognit.* 7, 95–112
69 Landauer, T.K. (1969) Reinforcement as consolidation. *Psychol. Rev.* 76, 82–96
70 Howard, M.W. and Kahana, M.J. (2002) A Distributed Representation of Temporal Context. *J. Math. Psychol.* 46, 269–299
71 Magill, R.A. and Hall, K.G. (1990) A review of the contextual interference effect in motor skill acquisition. *Hum. Mov. Sci.* 9, 241–289
72 Sinkeviciute, R. *et al.* (2019) The role of input variability and learner age in second language vocabulary learning. *Stud. Second Lang. Acquis.* 41, 795–820
73 Lev-Ari, S. (2018) The influence of social network size on speech perception. *Q. J. Exp. Psychol.* DOI: 10.1177/1747021817739865
74 Quinn, P.C. *et al.* (1993) Evidence for Representations of Perceptually Similar Natural Categories by 3-Month-Old and 4-Month-Old Infants. *Perception* 22, 463–475
75 Thibaut, J.-P. (1995) , The Abstraction of Relevant Features by Children and Adults: the Case of Visual Stimuli. , presented at the Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, 17, pp. 194
76 Kloos, H. and Sloutsky, V.M. (2008) What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *J. Exp. Psychol. Gen.* 137, 52–72
77 Hebert, E.P. *et al.* (1996) Practice Schedule Effects on the Performance and Learning of Low- and High-Skilled Students: An Applied Study. *Res. Q. Exerc. Sport* 67, 52–58
78 Twomey, K.E. *et al.* (2014) That's More Like It: Multiple Exemplars Facilitate Word Learning: Multiple Exemplars Facilitate Word Learning. *Infant Child Dev.* 23, 105–122
79 Heald, J.B. *et al.* (2020) Contextual inference underlies the learning of sensorimotor repertoires. *bioRxiv* DOI: 10.1101/2020.11.23.394320
80 Braithwaite, D.W. and Goldstone, R.L. (2015) Effects of Variation and Prior Knowledge on Abstract Concept Learning. *Cogn. Instr.* 33, 226–256

81 Gentner, D. and Hoyos, C. (2017) Analogy and Abstraction. *Top. Cogn. Sci.* 9, 672–693
82 Elman, J.L. (1993) Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99
83 Newport, E.L. (1990) Maturational Constraints on Language Learning. *Cogn. Sci.* 14, 11–28
84 Slone, L.K. *et al.* (2019) Self-generated variability in object images predicts vocabulary growth. *Dev. Sci.* 22, e12816
85 Weisberg, D.S. (2015) Pretend play. *WIREs Cogn. Sci.* 6, 249–261
86 Lillard, A.S. (2017) Why Do the Children (Pretend) Play? *Trends Cogn. Sci.* 21, 826–834
87 Sumner, M. (2011) The role of variation in the perception of accented speech. *Cognition* 119, 131–136
88 Carvalho, P.F. *et al.* (2021) The distributional properties of exemplars affect category learning and generalization. *Sci. Rep.* 11, 11263
89 Anderson, J.R. and Schooler, L.J. (1991) Reflections of the Environment in Memory. *Psychol. Sci.* 2, 396–408
90 Heederik, D. and von Mutius, E. (2012) Does diversity of environmental microbial exposure matter for the occurrence of allergy and asthma? *J. Allergy Clin. Immunol.* 130, 44–50
91 Wahlheim, C.N. *et al.* (2012) Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Mem. Cognit.* 40, 703–716
92 Fried, L.S. and Holyoak, K.J. (1984) Induction of category distributions: A framework for classification learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 234–257
93 Rips, L.J. (1989) Similarity, typicality, and categorization. *Similarity Analog. Reason.* 2159,
94 Markman, A.B. and Maddox, W.T. (2003) Classification of exemplars with single- and multiple-feature manifestations: The effects of relevant dimension variation and category structure. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 107–117
95 Mather, E. and Plunkett, K. (2011) Same items, different order: Effects of temporal variability on infant categorization. *Cognition* 119, 438–447
96 French, R.M. *et al.* (2004) The Role of Bottom-Up Processing in Perceptual Categorization by 3- to 4-Month-Old Infants: Simulations and Data. *J. Exp. Psychol. Gen.* 133, 382–397
97 Kornell, N. and Bjork, R.A. (2008) Learning Concepts and Categories: Is Spacing the “Enemy of Induction”? *Psychol. Sci.* 19, 585–592
98 Vlach, H.A. *et al.* (2012) At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. *J. Exp. Psychol. Learn. Mem. Cogn.* 38, 246–254
99 Oakes, L.M. and Ribar, R.J. (2005) A Comparison of Infants’ Categorization in Paired and Successive Presentation Familiarization Tasks. *Infancy* 7, 85–98
100 Kovack-Lesh, K.A. and Oakes, L.M. (2007) Hold your horses: How exposure to different items influences infant categorization. *J. Exp. Child Psychol.* 98, 69–93
101 Cepeda, N.J. *et al.* (2006) Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychol. Bull.* 132, 354–380
102 Moxley, S.E. (1979) Schema: The Variability of Practice Hypothesis. *J. Mot. Behav.* 11, 65–70
103 Schmidt, R.A. (1975) A schema theory of discrete motor skill learning. *Psychol. Rev.* 82, 225–260
104 Van Rossum, J.H.A. (1990) Schmidt’s schema theory: the empirical base of the variability of practice hypothesis. *Hum. Mov. Sci.* 9, 387–435
105 Desmottes, L. *et al.* (2017) Mirror-drawing skill in children with specific language impairment: Improving generalization by incorporating variability into the practice session. *Child Neuropsychol.* 23, 463–482
106 Adwan-Mansour, J. and Bitan, T. (2017) The Effect of Stimulus Variability on Learning and Generalization of Reading in a Novel Script. *J. Speech Lang. Hear. Res.* 60, 2840–2851
107 Arnold, G. and Auvray, M. (2018) Tactile recognition of visual stimuli: Specificity versus generalization of perceptual learning. *Vision Res.* 152, 40–50
108 Yao, W.X. *et al.* (2009) Variable Practice versus Constant Practice in the Acquisition of Wheelchair Propulsive Speeds. *Percept. Mot. Skills* DOI: 10.2466/pms.109.1.133-139
109 Braun, D.A. *et al.* (2009) Motor Task Variation Induces Structural Learning. *Curr. Biol.* 19, 352–357
110 Shea, J.B. and Morgan, R.L. (1979) Contextual interference effects on the acquisition, retention,

and transfer of a motor skill. *J. Exp. Psychol. [Hum. Learn.]* 5, 179–187
111 Barreiros, J. *et al.* (2007) The contextual interference effect in applied settings: *Eur. Phys. Educ. Rev.* DOI: 10.1177/1356336X07076876
112 Donovan, J.J. and Radosevich, D.J. (1999) A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *J. Appl. Psychol.* 84, 795–805
113 Lee, T.D. and Genovese, E.D. (1988) Distribution of Practice in Motor Skill Acquisition: Learning and Performance Effects Reconsidered. *Res. Q. Exerc. Sport* 59, 277–287
114 Porter, J.M. and Magill, R.A. (2010) Systematically increasing contextual interference is beneficial for learning sport skills. *J. Sports Sci.* 28, 1277–1285
115 Maye, J. *et al.* (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82, B101–B111
116 Singh, L. *et al.* (2004) Preference and processing: The role of speech affect in early spoken word recognition. *J. Mem. Lang.* 51, 173–189
117 Singh, L. *et al.* (2008) Building a Word-Form Lexicon in the Face of Variable Input: Influences of Pitch and Amplitude on Early Spoken Word Recognition. *Lang. Learn. Dev.* 4, 157–178
118 Sommers, M.S. and Barcroft, J. (2006) Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J. Acoust. Soc. Am.* 119, 2406–2416
119 Barriuso, T.A. and Hayes-Harb, R. (2018) High Variability Phonetic Training as a Bridge from Research to Practice. *CATESOL J.* 30, 177–194
120 Leong, C.X.R. *et al.* (2018) High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLOS ONE* 13, e0204888
121 Ingvalson, E.M. *et al.* (2014) Bilingual speech perception and learning: A review of recent trends. *Int. J. Biling.* 18, 35–47
122 Lively, S.E. *et al.* (1993) Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94, 1242–1255
123 Bradlow, A.R. and Bent, T. (2008) Perceptual adaptation to non-native speech. *Cognition* 106, 707–729
124 Estes, K.G. and Lew-Williams, C. (2015) Listening through voices: Infant statistical word segmentation across multiple speakers. *Dev. Psychol.* 51, 1517–1528
125 Atkinson, M. *et al.* (2015) Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. *PloS One* 10, e0129463
126 Adelman, J.S. *et al.* (2006) Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychol. Sci.*
127 Jones, M.N. *et al.* (2017) Context as an Organizing Principle of the Lexicon. In *Psychology of Learning and Motivation* 67pp. 239–283, Elsevier
128 Pagán, A. and Nation, K. (2019) Learning Words Via Reading: Contextual Diversity, Spacing, and Retrieval Effects in Adults. *Cogn. Sci.* 43, e12705
129 Hsiao, Y. and Nation, K. (2018) Semantic diversity, frequency and the development of lexical quality in children's word reading. *J. Mem. Lang.* 103, 114–126
130 Grunow, H. *et al.* (2006) The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *J. Commun. Disord.* 39, 158–170
131 Battig, W.F. (1972) Intratask interference as a source of facilitation in transfer and retention. *Top. Learn. Perform.*
132 Janiszewski, C. *et al.* (2003) A Meta-analysis of the Spacing Effect in Verbal Learning: Implications for Research on Advertising Repetition and Consumer Memory. *J. Consum. Res.* 30, 138–149
133 Glenberg, A.M. and Lehmann, T.S. (1980) Spacing repetitions over 1 week. *Mem. Cognit.* 8, 528–538
134 Dempster, F.N. (1987) Effects of variable encoding and spaced presentations on vocabulary learning. *J. Educ. Psychol.* 79, 162–170
135 Ellis, H.C. *et al.* (1974) Coding and varied input versus repetition in human memory. *J. Exp. Psychol.* 102, 284–290

136 Hernández-García, A. *et al.* (2018) , Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT. , in *2018 Conference on Cognitive Computational Neuroscience*, Philadelphia, Pennsylvania, USA
137 Mofid, N. *et al.* (2020) Keep Your AI-es on the Road: Tackling Distracted Driver Detection with Convolutional Neural Networks and Targeted Data Augmentation. *ArXiv200610955 Cs* at <http://arxiv.org/abs/2006.10955>
138 Perez, L. and Wang, J. (2017) The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *ArXiv171204621 Cs* at <http://arxiv.org/abs/1712.04621>
139 Park, D.S. *et al.* (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019* DOI: 10.21437/Interspeech.2019-2680
140 Shorten, C. and Khoshgoftaar, T.M. (2019) A survey on Image Data Augmentation for Deep Learning. *J. Big Data* 6, 60
141 Hernández-García, A. and König, P. (2018) Further advantages of data augmentation on convolutional neural networks. *ArXiv190611052 Cs* 11139, 95–103
142 Ragni, A. *et al.* (2014) , Data augmentation for low resource languages. , in *Procceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, pp. 810–814
143 Ko, T. *et al.* (2015) , Audio Augmentation for Speech Recognition. , in *Procceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, pp. 3586–3589
144 Cui, X. *et al.* (2015) Data Augmentation for Deep Neural Network Acoustic Modeling. *IEEEACM Trans. Audio Speech Lang. Process.* 23, 1469–1477
145 Schlüter, J. and Grill, T. (2015) EXPLORING DATA AUGMENTATION FOR IMPROVED SINGING VOICE DETECTION WITH NEURAL NETWORKS. *ISMIR*
146 Uhlich, S. *et al.* (2017) , Improving music source separation based on deep neural networks through data augmentation and network blending. , in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265
147 Carey, S. (1985) *Conceptual change in childhood*, MIT press.
148 Heit, E. and Hahn, U. (2001) Diversity-Based Reasoning in Children. *Cognit. Psychol.* 43, 243–273
149 Bjork, E.L. and Bjork, R.A. (2011) Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* pp. 56–64, Worth Publishers
150 Bjork, R.A. (1994) Memory and metamemory considerations in the training of human beings. *Metacognition Knowing Knowing*
151 Goldstone, R.L. (1994) *The role of Similarity in Categorization: Providing a Groundwork. Cognition*,
152 Hampton, J.A. (1998) Similarity-based categorization and fuzziness of natural categories. *Cognition* 65, 137–165
153 Kruschke, J.K. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychol. Rev.* 99, 22–44
154 Medin, D.L. and Schaffer, M.M. (1978) Context theory of classification learning. *Psychol. Rev.* 85, 207–238
155 Nosofsky, R.M. (1986) Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* 115, 39–57
156 Nosofsky, R.M. *et al.* (2019) Model-guided search for optimal natural-science-category training exemplars: A work in progress. *Psychon. Bull. Rev.* 26, 48–76
157 Hu, M. and Nosofsky, R.M. (2021) Exemplar-model account of categorization and recognition when training instances never repeat. *J. Exp. Psychol. Learn. Mem. Cogn.* DOI: 10.1037/xlm0001008
158 Smith, E.E. and Sloman, S.A. (1994) Similarity- versus rule-based categorization. *Mem. Cognit.* 22, 377–386
159 Thibaut, J.-P. *et al.* (2002) Dissociations between categorization and similarity judgments as a result of learning feature distributions. *Mem. Cognit.* 30, 647–656
160 Shepard, R. (1987) Toward a universal law of generalization for psychological science. *Science*

237, 1317–1323
161 Pavlik, P.I. and Anderson, J.R. (2005) Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cogn. Sci.* 29, 559–586
162 Wickelgren, W.A. (1972) Trace resistance and the decay of long-term memory. *J. Math. Psychol.* 9, 418–455
163 Hintzman, D.L. (1974) Theoretical implications of the spacing effect.
164 Johnson, K. (1997) Speech perception without speaker normalization: An exemplar model. In *Talker Variability in Speech Processing* pp. 145–165, Academic Press
165 Elman, J.L. (1990) Finding Structure in Time. *Cogn. Sci.* 14, 179–211
166 Christiansen, M.H. *et al.* (1998) Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Lang. Cogn. Process.* 13, 221–268
167 Hummel, J.E. and Holyoak, K.J. (2003) A symbolic-connectionist theory of relational inference and generalization. *Psychol. Rev.* 110, 220–264
168 McClelland, J.L. *et al.* (1986) *Parallel distributed processing: explorations in the microstructure, vol. 2: psychological and biological models*, 2MIT Press.
169 Krizhevsky, A. *et al.* (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 25, 1097–1105
170 Taigman, Y. *et al.* (2014) , DeepFace: Closing the Gap to Human-Level Performance in Face Verification. , in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708
171 Collobert, R. *et al.* (2011) Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* 12, 2493–2537
172 Silver, D. *et al.* (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 1140–1144
173 Hornik, K. *et al.* (1989) Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366
174 Zhu, X. (2015) , Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education. , in *AAAI*
175 Beale, W.H. (2009) *Learning from Language*, University of Pittsburgh Press.
176 Mitroff, S.R. *et al.* (2013) Enhancing Ice Hockey Skills Through Stroboscopic Visual Training: A Pilot Study. *Athl. Train. Sports Health Care* 5, 261–264
177 Smith, T.Q. and Mitroff, S.R. (2012) Stroboscopic Training Enhances Anticipatory Timing. *Int. J. Exerc. Sci.* 5, 344–353
178 Tauber, S.K. *et al.* (2013) Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychon. Bull. Rev.* 20, 356–363
179 Allcott, H. *et al.* (2007) Community Size and Network Closure. *Am. Econ. Rev.* 97, 80–85
180 Bahlmann, M.D. (2014) Geographic Network Diversity: How Does it Affect Exploratory Innovation? *Ind. Innov.* 21, 633–654
181 Liu, B.S.-C. *et al.* (2005) DiffuNET: The impact of network structure on diffusion of innovation. *Eur. J. Innov. Manag.* DOI: 10.1108/14601060510594701
182 Balas, B. and Saville, A. (2015) N170 face specificity and face memory depend on hometown size. *Neuropsychologia* 69, 211–217
183 Lev-Ari, S. and Shao, Z. (2016) How social network heterogeneity facilitates lexical access and lexical prediction. *Mem. Cognit.* DOI: 10.3758/s13421-016-0675-y
184 Raviv, L. *et al.* (2019) Larger communities create more systematic languages. *Proc. R. Soc. B Biol. Sci.* 286, 20191262
185 Stephan, G.E. and McMullin, D.R. (1982) Tolerance of Sexual Nonconformity: City Size as a Situational and Early Learning Determinant. *Am. Sociol. Rev.* 47, 411–415
186 Park, B. *et al.* (1991) Social Categorization and the Representation of Variability Information. *Eur. Rev. Soc. Psychol.* 2, 211–245
187 Cinelli, M. *et al.* (2021) The echo chamber effect on social media. *Proc. Natl. Acad. Sci.* 118, e2023301118
188 Sunstein, C.R. (1999) *The Law of Group Polarization*, Social Science Research Network.