Introduction

immediate

Author note

Introduction

Introduction

The past century of research on human learning has produced ample evidence that although learners can improve at almost any task, such improvements are often specific to the trained task, with unreliable or even nonexistent transfer or generalization to novel tasks or conditions (Barnett & Ceci, 2002; Detterman, 1993). Such generalization challenges are of noteworthy practical relevance, given that educators, trainers, and rehabilitators typically intend for their students to be able to apply what they have learned to new situations. It is therefore important to better understand the factors that influence generalization, and to develop cognitive models that can predict when generalization is likely to occur. Such characteristics have included training difficulty, spacing, temporal order, feedback schedules, and the primary focus of the current work - the variability of training examples.

The study of variability

Varied training has been shown to influence learning in wide array of different tasks and domains, including categorization (Hahn et al., 2005; Maddox & Filoteo, 2011; Morgenstern et al., 2019; Nosofsky et al., 2019; Plebanek & James, 2021; Posner & Keele, 1968), language learning (Brekelmans et al., 2022; Jones & Brandt, 2020; Perry et al., 2010; Twomey et al., 2018; Wonnacott et al., 2012), anagram completion (Goode et al., 2008), trajectory extrapolation (Fulvio et al., 2014), cognitive control tasks (Moshon-Cohen et al., 2024; Sabah et al., 2019), associative learning (Fan et al., 2022; Lee et al., 2019; Prada & Garcia-Marques, 2020; Reichmann et al., 2023), visual search (George & Egner, 2021; Gonzalez & Madhavan, 2011; Kelley & Yantis, 2009), voice identity learning

(Lavan et al., 2019), face recognition (Burton et al., 2016; Honig et al., 2022; Menon et al., 2015), the perception of social group heterogeneity (Gershman & Cikara, 2023; Konovalova & Le Mens, 2020; Linville & Fischer, 1993; Park & Hastie, 1987) , simple motor learning (Braun et al., 2009; Kerr & Booth, 1978; Roller et al., 2001; Willey & Liu, 2018a), sports training (Breslin et al., 2012; Green et al., 1995; North et al., 2019), and complex skill learning (Hacques et al., 2022; Huet et al., 2011; Seow et al., 2019). See Czyż (2021) or Raviv et al. (2022) for more detailed reviews.

Research on the effects of varied training typically manipulates variability in one of two ways. In the first approach, a high variability group is exposed to a greater number of unique instances during training, while a low variability group receives fewer unique instances with more repetitions. Alternatively, both groups may receive the same number of unique instances, but the high variability group's instances are more widely distributed or spread out in the relevant psychological space, while the low variability group's instances are clustered more tightly together. Researchers then compare the training groups in terms of their performance during the training phase, as well as their generalization performance during a testing phase. Researchers will usually compare the performance of the two groups during both the training phase and a subsequent testing phase. The primary theoretical interest is often to assess the influence of training variability on generalization to novel testing items or conditions. However, the test may also include some or all of the items that were used during the training stage, allowing for an assessment of whether the variability manipulation influenced the learning of the trained items themselves, or to make it easy to measure how much performance degrades as a function of how far away testing items are from the training items.

The influence of training variation has received a large amount of attention in the domain of sensorimotor skill learning. Much of this research has been influenced by the work of Schmidt

(1975), who proposed a schema-based account of motor learning as an attempt to address the longstanding problem of how novel movements are produced. Schema theory presumes that learners possess general motor programs for a class of movements (e.g. an underhand throw). When called up for use motor programs are parameterized by schema rules which determine how the motor program is parameterized or scaled to the particular demands of the current task. Schema theory predicts that variable training facilitates the formation of more robust schemas, which will result in improved generalization or transfer. Experiments that test this hypothesis are often designed to compare the transfer performance of a constant-trained group against that of a varied-trained group. Both groups train on the same task, but the varied group practices with multiple instances along some task-relevant dimension that remains invariant for the constant group. For example, studies using a projectile throwing task might assign participants to either constant training that practicing throwing from a single location, and a varied group that throws from multiple locations. Following training, both groups are then tested from novel throwing locations (Pacheco & Newell, 2018; Pigott & Shapiro, 1984; Willey & Liu, 2018a; Wulf, 1991).

One of the earliest, and still often cited investigations of Schmidt's benefits of variability hypothesis was the work of Kerr and Booth (1978). Two groups of children, aged 8 and 12, were assigned to either constant or varied training of a bean bag throwing task. The constant group practiced throwing a bean-bag at a small target placed 3 feet in front of them, and the varied group practiced throwing from a distance of both 2 feet and 4 feet. Participants were blindfolded and unable to see the target while making each throw but would receive feedback by looking at where the beanbag had landed in between each training trial. 12 weeks later, all of the children were given a final test from a distance of 3 feet which was novel for the varied participants and repeated for the constant participants. Participants were also blindfolded for testing and did not receive

trial by trial feedback in this stage. In both age groups, participants performed significantly better in the varied condition than the constant condition, though the effect was larger for the younger, 8-year-old children. This result offers a particularly compelling example of the merits of varied practice, given that the varied group was able to outperform the constant group even from the home turf location where one may have expected the constatn group to have the strongest advantage. A similar pattern of results was observed in another study wherein varied participants trained with tennis, squash, badminton, and short-tennis rackets were compared against constant subjects trained with only a tennis racket (Green et al., 1995). One of the testing conditions had subjects repeat the use of the tennis racket, which had been used on all 128 training trials for the constant group, and only 32 training trials for the varied group. Nevertheless, the varied group outperformed the constant group when using the tennis racket at testing, and also performed better in conditions with several novel racket lengths. However, as is the case with many of the patterns commonly observed in the "benefits of variability" literature, the pattern wherein the varied group outperfroms the constant group even from the constants group's home turf has not been consistently replicated. One recent study attempted a near replication of the Kerr & Booth study (Willey & Liu, 2018b), having subjects throw beanbags at a target, with the varied group training from positions (5 and 9 feet) on either side of the constant group (7 feet). This study did not find a varied advantage from the constant training position, though the varied group did perform better at distances novel to both groups. However, this study diverged from the original in that the participants were adults; and the amount of training was much greater (20 sessions with 60 practice trials each, spread out over 5-7 weeks).

Pitting varied against constant practice against each other on the home turf of the constant group provides a compelling argument for the benefits of varied training, as well as an interesting

challenge for theoretical accounts that posit generalization to occur as some function of distance. However, despite its appeal this particular contrast is relatively uncommon in the literature. It is unclear whether this may be cause for concern over publication bias, or just researchers feeling the design is too risky. A far more common design is to have separate constant groups that each train exclusively from each of the conditions that the varied group encounters (Catalano & Kleiner, 1984; Chua et al., 2019; McCracken & Stelmach, 1977; Moxley, 1979; Newell & Shapiro, 1976), or for a single constant group to train from just one of the conditions experienced by the varied participants (Pigott & Shapiro, 1984; Roller et al., 2001; Wrisberg & McLean, 1984; Wrisberg & Mead, 1983). A less common contrast places the constant group training in a region of the task space outside of the range of examples experienced by the varied group, but distinct from the transfer condition (Wrisberg et al., 1987; Wulf & Schmidt, 1997). Of particular relevance to the current work is the early study of Catalano and Kleiner (1984), as theirs was one of the earliest studies to investigate the influence of varied vs. constant training on multiple testing locations of graded distance from the training condition. Participants were trained on coincident timing task, in which subjects observe a series of lightbulbs turning on sequentially at a consistent rate and attempt to time a button response with the onset of the final bulb. The constant groups trained with a single velocity of either 5,7,9, or 11 mph, while the varied group trained from all 4 of these velocities. Participants were then assigned to one of four possible generalization conditions, all of which fell outside of the range of the varied training conditions – 1, 3, 13 or 15 mph. As is often the case, the varied group performed worse during the training phase. In the testing phase, the general pattern was for all participants to perform worse as the testing conditions became further away from the training conditions, but since the drop off in performance as a function of distance was far less steep for the varied group, the authors suggested that varied training induced

a decremented generalization gradient, such that the varied participants were less affected by the change between training and testing conditions.

Benefits of varied training have also been observed in many studies outside of the sensorimotor domain. Goode et al. (2008) trained participants to solve anagrams of 40 different words ranging in length from 5 to 11 letters, with an anagram of each word repeated 3 times throughout training, for a total of 120 training trials. Although subjects in all conditions were exposed to the same 40 unique words (i.e. the solution to an anagram), participants in the varied group saw 3 different arrangements for each solution-word, such as DOLOF, FOLOD, and OOFLD for the solution word FLOOD, whereas constant subjects would train on three repetitions of LDOOF (spread evenly across training). Two different constant groups were used. Both constant groups trained with three repetitions of the same word scramble, but for constant group A, the testing phase consisted of the identical letter arrangement to that seen during training (e.g. LDOOF), whereas for constant group B, the testing phase consisted of a arrangement they had not seen during training, thus presenting them with a testing situation similar situation to the varied group. At the testing stage, the varied group outperformed both constant groups, a particularly impressive result, given that constant group A had 3 prior exposures to the word arrangement (i.e. the particular permutation of letters) which the varied group had not explicitly seen. However varied subjects in this study did not exhibit the typical decrement in the training phase typical of other varied manipulations in the literature, and actually achieved higher levels of anagram solving accuracy by the end of training than either of the constant groups – solving 2 more anagrams on average than the constant group. This might suggest that for tasks of this nature where the learner can simply get stuck with a particular word scramble, repeated exposure to the identical scramble might be less helpful towards finding the solution than being given a different arrangement of

the same letters. This contention is supported by the fact that constant group A, who was tested on the identical arrangement as they experienced during training, performed no better at testing than did constant group B, who had trained on a different arrangement of the same word solution – further suggesting that there may not have been a strong identity advantage in this task.

In the domain of category learning, the constant vs. varied comparison is much less suitable. Instead, researchers will typically employ designs where all training groups encounter numerous stimuli, but one group experiences a greater number of unique exemplars (Brunstein & Gonzalez, 2011; Doyle & Hourihan, 2016; Hosch et al., 2023; Nosofsky et al., 2019; Wahlheim et al., 2012), or designs where the number of unique training exemplars is held constant, but one group trains with items that are more dispersed, or spread out across the category space (Bowman & Zeithamova, 2020; Homa & Vosburgh, 1976; Hu & Nosofsky, 2024; Maddox & Filoteo, 2011; Posner & Keele, 1968).

Much of the earlier work in this sub-area trained subjects on artificial categories, such as dot patterns (Homa & Vosburgh, 1976; Posner & Keele, 1968). A seminal study by Posner and Keele (1968) trained participants to categorize artificial dot patterns, manipulating whether learners were trained with low variability examples clustered close to the category prototypes (i.e. low distortion training patterns), or higher-variability patterns spread further away from the prototype (i.e. high-distortion patterns). Participants that received training on more highly-distorted items showed superior generalization to novel high distortion patterns in the subsequent testing phase. It should be noted that unlike the sensorimotor studies discussed earlier, the Posner and Keele (1968) study did not present low-varied and high-varied participants with an equal number of training rathers, but instead had participants remain in the training stage of the experiment until they reached a criterion level of performance. This train-until-criterion procedure led to

the high-variability condition participants tending to complete a larger number of training trials before switching to the testing stage. More recent work (Hu & Nosofsky, 2024), also used dot pattern categories, but matched the number of training trials across conditions. Under this procedure, higher-variability participants tended to reach lower levels of performance by the end of the training stage. The results in the testing phase were the opposite of Posner and Keele (1968), with the low-variability training group showing superior generalization to novel high-distortion patterns (as well as generalization to novel patterns of low or medium distortion levels). However, whether this discrepancy is solely a result of the different training procedures is unclear, as the studies also differed in the nature of the prototype patterns used. Posner and Keele (1968) utilized simpler, recognizable prototypes (e.g., a triangle, the letter M, the letter F), while Hu and Nosofsky (2024) employed random prototype patterns.

Recent studies have also begun utilizing more complex or realistic sitmuli when assessing the influence of variability on category learning. Wahlheim et al. (2012) conducted one such study. In a within-participants design, participants were trained on bird categories with either high repetitions of a few exemplars, or few repetitions of many exemplars. Across four different experiments, which were conducted to address an unrelated question on metacognitive judgements, the researchers consistently found that participants generalized better to novel species following training with more unique exemplars (i.e. higher variability), while high repetition training produced significantly better performance categorizing the specific species they had trained on. A variability advantage was also found in the relatively complex domain of rock categorization (Nosofsky et al., 2019). For 10 different rock categories, participants were trained with either many repetitions of 3 unique examples of each category, or few repetitions of 9 unique examples, with an equal number of total training trials in each group (the design also included 2 other con-