Journal of Memory and Language 138 (2024) 104531

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Modality and stimulus effects on distributional statistical learning: Sound vs. sight, time vs. space

Haoyu Zhou [a,*], Sabine van der Ham [b], Bart de Boer [c], Louisa Bogaerts [a,1], Limor Raviv [c,d,e,1]

[a] Ghent University, Belgium
[b] Hanzehogeschool Groningen, the Netherlands
[c] Vrije Universiteit Brussel AI-Lab, Belgium
[d] LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[e] cSCAN, University of Glasgow, Glasgow, UK

ARTICLE INFO

Keywords:
Statistical Learning
Distributional learning
Frequency distribution
Modality-specificity
Audition vs. Vision
Categories

ABSTRACT

Statistical learning (SL) is postulated to play an important role in the process of language acquisition as well as in other cognitive functions. It was found to enable learning of various types of statistical patterns across different sensory modalities. However, few studies have distinguished distributional SL (DSL) from sequential and spatial SL, or examined DSL across modalities using comparable tasks. Considering the relevance of such findings to the nature of SL, the current study investigated the modality- and stimulus-specificity of DSL. Using a within-subject design we compared DSL performance in auditory and visual modalities. For each sensory modality, two stimulus types were used: linguistic versus non-linguistic auditory stimuli and temporal versus spatial visual stimuli. In each condition, participants were exposed to stimuli that varied in their length as they were drawn from two categories (short versus long). DSL was assessed using a categorization task and a production task. Results showed that learners' performance was only correlated for tasks in the same sensory modality. Moreover, participants were better at categorizing the temporal signals in the auditory conditions than in the visual condition, where in turn an advantage of the spatial condition was observed. In the production task participants exaggerated signal length more for linguistic signals than non-linguistic signals. Together, these findings suggest that DSL is modality- and stimulus-sensitive.

Introduction

Statistical learning

Statistical learning (SL) is a powerful cognitive tool to extract statistical regularities from sensory input, which enables learners to detect structure in the vast amounts of sensory information they are exposed to. This sensitivity to statistical regularities has been investigated extensively in the field of cognitive science (Bogaerts, Frost, & Christiansen, 2020) as well as in linguistics given the important role it is postulated to play in the process of language acquisition (Romberg & Saffran, 2010; Saffran, 2003).

Conditional vs. distributional SL

Two types of input regularities can be distinguished for SL (Siegelman et al., 2017). The first type pertains to sequential and spatial relations between stimuli, such as the co-occurrence of particular sounds and shapes in time (e.g., Saffran, Aslin, & Newport, 1996; Turk-Browne, Jungé, & Scholl, 2005) or in space (e.g., Fiser & Aslin, 2001; Orbán, Fiser, Aslin, & Lengyel, 2008). Because this type of SL refers to the learning of regularities such as conditional probabilities, we refer to it as conditional statistical learning (CSL, see also Thiessen & Erickson, 2013; Growns, Siegelman, & Martire, 2020). As CSL helps to identify how sequences or complex scenes are formed from a set of discrete building blocks, it is often investigated in the context of speech segmentation and syntactic processing (Conway, Bauernschmidt, Huang, & Pisoni, 2010; Misyak & Christiansen, 2012) as well as in the context of visual

* Corresponding author at: Henri Dunantlaan 2, 9000 Ghent, Belgium.
E-mail addresses: haoyu.zhou@ugent.be (H. Zhou), a.j.s.van.der.ham@pl.hanze.nl (S. van der Ham), bart@ai.vub.ac.be (B. de Boer), louisa.bogaerts@ugent.be (L. Bogaerts), limor.raviv@mail.huji.ac.il (L. Raviv).
[1] Equal contributions as last authors.

https://doi.org/10.1016/j.jml.2024.104531
Received 5 September 2023; Received in revised form 11 April 2024; Accepted 12 April 2024
Available online 30 April 2024

processing (Fiser & Lengyel, 2022). The second type of regularity pertains to the frequency distribution of individual exemplars. Distributional statistical learning (DSL) has mainly been investigated in the context of phonology and category learning (Maye et al., 2002, 2008).

CSL and DSL are proposed to be supported by distinct yet interrelated memory processes and differ in the (implicit) knowledge acquired (Thiessen & Erickson, 2013). When facing a continuous sequential input, CSL relies on transitional probabilities to index sequential relations within the input stream and extract novel discrete representations of repeating patterns (Thiessen et al., 2013). For example, the seminal study by Saffran and colleagues (1996) revealed that infants as young as 8 months old can track transitional probability information from an artificial speech stream of auditory syllables, allowing them to extract repeated words without any additional cues for word boundaries and to showcase familiarity with these words. In similar experiments as Saffran et al. (1996), this type of SL was found to facilitate the learning of various linguistic structures, including word order (Gervain, Macagno, Cogoi, Peña, & Mehler, 2008), syntactic patterns (Gomez & Gerken, 1999), and phonotactics (Chambers, Onishi, & Fisher, 2003). Outside the linguistic domain, the sensitivity to conditional relationships in sequential input was also demonstrated with visual (Kirkham, Slemmer, & Johnson, 2002; Zimmerer, Cowell, & Varley, 2010) and tactile stimuli (Conway & Christiansen, 2005; Conway & Christiansen, 2006). In the context of spatial input, sensitivity to conditional regularities was identified with visual stimuli featuring spatial configurations where multiple elements were presented simultaneously (e.g., Fiser & Aslin, 2001; Orbán et al., 2008). Moreover, CSL was documented across a wide range of age groups (Raviv & Arnon, 2018; Saffran, Johnson, Aslin, & Newport, 1999) and even in non-human species (Milne, Petkov, & Wilson, 2018; Sonnweber, Ravignani, & Fitch, 2015). Taken together, these results show that the sensitivity to conditional regularities in the input is present not only in human language learning but also in other domains.

By contrast, information regarding the frequency, variance, and context of multiple exemplars is aggregated during DSL. Thiessen et al. (2013) proposed that, unlike CSL, where the extracted patterns form discrete representations in long-term memory, it is the integration across such discrete representations that gives rise to learners' sensitivity to the distribution underlying the input and the discovery of categorical structure. For instance, Maye and colleagues (2002) showed that infants can categorize speech sounds taken from a phonetic continuum according to the bimodal frequency distribution of the input. After a familiarization phase with an input stream that either contained a bimodal or a unimodal frequency distribution of tokens from a [ta] to [da] continuum, only infants who were exposed to the bimodal distribution successfully discriminated tokens from the endpoints of the continuum. This result indicated that infants use distributional information to make sense of the acoustic variability that characterizes speech and learn the underlying phonetic structure of the language. Extending the results of Maye, Werker, and Gerken (2002), learning of the statistical distribution of sounds for forming phoneme categories was also documented in children (Vandermosten, Wouters, Ghesquière, & Golestani, 2019; see Cristia, 2018 for a review), adults (Hayes-Harb, 2007; Maye & Gerken, 2011), and non-human species (Pons, 2006). Considering other domains and types of input, DSL has been demonstrated with discrimination tasks of non-native lexical tones (Liu et al., 2022), musical pitches (Ong, Burnham, & Stevens, 2017), as well as shapes that differ in size (Rosenthal, Fusi, & Hochstein, 2001) and human faces (Altvater-Mackensen, Jessen, & Grossmann, 2017). In addition, recent investigations on the effect of statistical regularities on visual selection have found that participants give attentional priority to locations in a visual display where targets are likely to appear and suppress locations where distractors appear with higher probability (Theeuwes, Bogaerts, & van Moorselaar, 2022). This finding shows that DSL can also contribute to optimizing attention allocation and visual processing.

Although distributional patterns are a major component of language learning and general pattern detection, they have received much less attention. This is clearly illustrated in the review by Frost and colleagues (2019), which pointed out that the vast majority of SL studies have focused on sequential conditional regularities, using paradigms with embedded triplets and pairs akin to Saffran et al. (1996)'s seminal study, or artificial grammar learning (Reber, 1969) which is commonly used in the implicit learning literature yet arguably measures a similar type of learning (Perruchet & Pacton, 2006; Christiansen, 2019). The focus of the current investigation is the modality- and stimulus-sensitivity of learning distributional regularities, and more specifically the learning of categories based on signals that vary in their length. In what follows we discuss previous research on the constraints on statistical learning at large, including CSL, as most works focused on the learning of sequential regularities and these findings set the stage for the current study and shaped our predictions.

Constraints on SL

DSL and CSL have both been demonstrated across sensory modalities (e.g., Auditory vs. Visual) and across stimulus types (Linguistic vs. Non-linguistic). Meanwhile, multiple studies that examined learning across different domains or modalities showed limited transfer, minimal interference or low correlation and thus reveal modality- and stimulus-specificity. This leads to the question of whether SL is a domain-general or domain-specific learning mechanism. The view proposed by Frost, Armstrong, Siegelman, and Christiansen (2015) suggests that SL is a set of domain-general principles for learning that are nevertheless subject to modality and stimulus constraints given the specific characteristics of the brain regions that are involved. We will discuss evidence for these constraints in more detail below. What is worth noting in preview is that much of what we know about SL as a general ability or learning principle comes from studies that specifically target CSL with sequential input, with only little known on how DSL adheres to these general patterns (e.g., Maye et al., 2002; Thiessen, 2011).

Modality-sensitivity in CSL

Initial work that investigated whether SL is bound by modality constraints revealed a mixed pattern of results for sequential regularities. While some research reported different learning outcomes across modalities (e.g., Conway & Christiansen, 2005; Conway & Christiansen, 2009; Milne et al., 2018), others found no clear learning advantage in either the visual or auditory modality (e.g., Zimmerer et al., 2010). Today, there is growing evidence that supports modality-sensitive models of SL (e.g., Frost et al., 2015; Krogh, Vlach, & Johnson, 2012; Pavlidou & Bogaerts, 2019; Silva, Folia, Inácio, Castro, & Petersson, 2018).

One of the first studies that asked whether CSL operates differently across domains was performed by Conway and Christiansen (2005), who directly compared performance in an artificial grammar learning task in the auditory, visual, and tactile modalities. Participants were exposed to a series of sequences generated according to an underlying grammar and subsequently had to judge the grammaticality of novel sequences. Despite the fact that all conditions had the same underlying structure, results showed that performance was better in the auditory condition versus the visual and tactile conditions. Additionally, there were qualitative differences between modalities: participants in the tactile group were more sensitive to sequence-initial information, while participants in the auditory group showed sensitivity to sequence-final information. These findings suggest that sensitivity to sequential conditional regularities in a given sensory input is subject to modality constraints. These constraints have been underscored by later studies, such as one that made use of an embedded pattern paradigm which showed that visual and auditory CSL are affected differently by presentation rates: while a faster presentation rate improves learning in the auditory domain, it hinders learning in the visual domain (Emberson, Conway, &

Christiansen, 2011). This divergent timing effect is in line with results from perceptual studies, which suggest that the auditory modality lends itself better to processing rapid temporal information compared to the visual modality (e.g., see Grondin, 2010 for review).

While the above studies show that sensitivity to sequential regularities in the sensory input is not uniform across modalities, their between-subjects design did not allow for comparison at an individual level. What remains unclear is whether individuals' sensitivity to statistical patterns in one modality is predictive of their sensitivity to similar patterns in another modality. In other words, are some individuals simply better than others at detecting statistical patterns across the board, or are these abilities in different sensory modalities uncorrelated even within an individual? To address this question, Siegelman and Frost (2015) tested subjects with a task battery of four CSL tasks that differed in the sensory modality, with stimuli that were either verbal or nonverbal, and sequential statistical contingencies that either spanned adjacent elements or concerned non-adjacent elements (i.e., auditory-verbal-adjacent; auditory-verbal-nonadjacent; auditory-nonverbal-adjacent; visual-nonverbal-adjacent). Interestingly, they found evidence for stable individual differences in a given task, yet no correlation between individuals' performance on the different CSL tasks was observed. The researchers concluded that SL is not a unified capacity: people's sensitivity to sequential statistical patterns seems to be determined by the sensory modality and by the specific stimulus type (see also Siegelman et al., 2017). Furthering this componential view, Bogaerts and colleagues (2022) proposed that there might not even be such a thing as a "good statistical learner" in the absence of a general SL capacity that can sort individuals from bad to good learners.

Although growing evidence points to modality-based differences in SL abilities, it is hard to draw clear conclusions regarding the generality of these findings for several reasons. First, only a few studies have compared SL performance across visual and auditory (unimodal) conditions using a within-individuals design (Pavlidou & Bogaerts, 2019; Siegelman & Frost, 2015). Second, and perhaps most important, no study to date has examined modality-based differences in DSL. A series of studies have shown multimodal facilitation in distributional learning (e.g., Mani & Schneider, 2013; Mitchel, Gerfen, & Weiss, 2016; Teinonen, Aslin, Alku, & Csibra, 2008), but the effect of different modalities on the learning outcomes is yet to be explored. Considering that more and more studies are showing direct evidence of DSL for either auditory or visual categorization (e.g., Broedelet, Boersma, & Rispens, 2022; Jung, Walther, & Finn, 2021), the next step is to directly compare DSL in different modalities. The current study will do just that, but in addition investigates whether the specific nature of stimuli within a given sensory modality impacts learning.

Stimulus-sensitivity

A linguistic advantage? Does SL with linguistic input work differently than with other stimuli? Given that other species are also able to pick up statistical regularities, it is unlikely that SL has originally evolved for language learning (see review by Santolin & Saffran, 2018). Regardless, previous studies have shown that individual differences in SL are predictive of a wide range of language-related outcomes in both children (Arciuli & Simpson, 2012; Kidd, 2012; Shafto, Conway, Field, & Houston, 2012) and adults (Conway et al., 2010; Misyak & Christiansen, 2012). Moreover, it was shown that unlike visual CSL (and most other cognitive abilities which improve with age), auditory CSL seems to be age-invariant and does not change much across childhood (Raviv & Arnon, 2018). The language-specificity of SL could be due to the auditory input also being linguistic in nature (see also Boeve, Zhou, & Bogaerts, in press). Indeed, a follow-up study indicated that the developmental trajectories of visual and non-linguistic auditory SL are in fact similar when the auditory SL task used familiar sounds (e.g., a bird tweeting, a door opening) rather than syllables (Shufaniya & Arnon, 2018). Learning linguistic materials and non-linguistic materials were also found to be differentially affected by concurrent motor production – the learning of linguistic sequences was hindered when participants had to whisper, whereas the learning of non-linguistic sequences was not (Boeve, Möttönen, & Smalle, 2024). Together, these findings suggest that linguistic information may be processed and learned differently compared to non-linguistic information. A likely explanation lies in human learners' extensive previous exposure to sequences of linguistic auditory stimuli (e.g., syllables in speech; Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018).

Although learners' prior knowledge about linguistic structure can affect SL performance both positively and negatively depending on the (mis)match of the to-be-learned regularities with participants' prior knowledge (Elazar et al., 2022; Siegelman et al., 2018), evidence for a learning advantage with linguistic stimuli was found in the recent study of Lukics and Lukács (2022) that reported an overall advantage of auditory-linguistic stimuli compared to other types of stimuli (including visual-linguistic input). Notably, the effect of linguistic versus non-linguistic materials has only been investigated for sequential conditional regularities, and there is currently no data on whether DSL shows a similar advantage for linguistic stimuli.

Stimulus familiarity. Moving beyond differences in learning with linguistic and non-linguistic stimuli, Perfors and Kidd (2022) investigated visual SL performance as a function of how familiar the stimuli were to learners. Perceptual fluency, capturing how efficiently people can encode the individual stimuli, was shown to be strongly driven by stimulus familiarity (whereas it was independent of stimulus complexity) and was positively correlated with the CSL performance. Considering that the above findings centered on a language-specificity effect on auditory CSL due to participants' prior knowledge, the familiarity of the stimuli seems to affect learners' performance across modalities.

Time versus space. Finally, visual CSL is also sensitive to whether the input is temporal (i.e., sequential) versus spatial. For example, when contrasting learning of statistically governed input that was either presented temporally (i.e., color squares appearing sequentially in the center of the screen) or spatially (i.e., the same color squares appearing simultaneously in four locations along a horizontal row), it was found that participants showed better learning in the spatial presentation format (Conway & Christiansen, 2009). Moreover, a faster stimulus presentation rate negatively affected performance only in the temporal task.

The contrast between temporal versus spatial input has also been investigated in a recent study that directly compared participants' performance in CSL and DSL. Growns and colleagues (2020) conducted four SL tasks, testing the learning of conditional regularities and distributional regularities across time and space. Conditional regularities in time were operationalized as co-occurrences between centrally presented shapes, whereas spatial co-occurrences were pairs of shapes appearing with a predetermined spatial relationship in a grid. In the distributional version of the tasks, shapes were presented at the 'arms' of a series of snowflakes and each shape appeared with a different frequency. In the nonspatial task, appearances were evenly distributed across all arms, whereas in the spatial version each shape appeared with different frequencies in different locations. In all tasks, learning was measured with a combination of pattern recognition and pattern completion trials testing the preference for a correct statistical pattern over a foil pattern. Significant but moderate correlations were observed among participants' performance on all four tasks, and results of a principal component analysis showed that a large proportion of the variance in performance was explained by a shared component. Interestingly, a smaller but substantial portion of the variance was accounted for by performance on each of the individual tasks. These findings suggest that visual SL of conditional and/or distributional regularities is the result of the interplay between a unified learning mechanism and individuals' ability to encode specific input and extract specific types of regularities.

In sum, a handful of studies so far have documented modality- and stimulus-specificity by examining learning in different sensory

modalities, and with different types of stimuli, including linguistic vs. non-linguistic, and temporal vs. spatial stimuli (see also Frost et al., 2015, for review). However, the near-exclusive focus on CSL of sequential regularities leaves open the question whether modality- and stimulus-sensitivity also characterize DSL. In addition, previous works targeted the learning of patterns involving various stimulus identities, presented either sequentially or in spatial configurations, leaving unexplored modality- and stimulus-effect in how people learn to distinguish between different versions of the same stimulus.

The current study

The current preregistered study focuses exclusively on DSL, testing learners' ability to extract categorical information from continuous input based on its distribution and examining the modality-sensitive and stimulus-sensitive nature of this type of learning. The full preregistration can be found here: https://osf.io/34agt. More specifically, we ask:

1. Is distributional SL a *modality*-sensitive ability?
1. Is distributional SL a *stimulus*-sensitive ability?

Based on previous studies of SL reviewed above, which showed that there are important differences between the visual and the auditory modalities (e.g., Emberson et al., 2011), linguistic and non-linguistic auditory stimuli (e.g., Siegelman et al., 2018), and temporal vs. spatial visual stimuli (e.g., Conway & Christiansen, 2009), we conducted a within-subject comparison of DSL abilities in four experimental conditions, spanning different modalities and stimuli: Auditory-Linguistic condition, Auditory-Non-Linguistic condition, Visual-Temporal condition, and Visual-Spatial condition. This design will allow us to compare participants' performance between sensory modalities (i.e., auditory versus visual) and between stimulus types within a modality (i.e., linguistic versus non-linguistic stimuli in the auditory domain, and temporal versus spatial stimuli in the visual domain). In all tasks participants learned about the length distributions of two categories of stimuli: short versus long. In temporal conditions, this length corresponded to the duration of stimuli. In the Visual-Spatial condition length corresponded to the height of a stimulus (see ***Materials and Stimuli*** for details).

In brief, each experimental condition consisted of three parts (see ***Procedure***): (1) an exposure phase, (2) a categorization task, and (3) a production task. During the exposure phase, participants were exposed to signals from a bimodal distribution that varied in their duration or their height, with each signal being associated with one of two categories. Specifically, short and long signals corresponded to either "dangerous" or "safe" aliens. To assess participants' knowledge of their input, participants were asked to categorize signals as well as produce typical signals of the categories they have learned, following van der Ham and de Boer (2015). In the categorization task, participants were exposed to all signals one by one, and for each signal they needed to indicate which category the signal belongs to. Their categorization accuracy was then measured. In the production task, participants were asked to produce representative signals for each category using the computer interface, and their reproduction accuracy was measured by examining how much their productions deviated from their input (i.e., how accurately participants reproduced a representative signal for a given category). This is a unique design, considering that most SL studies only include an alternative forced-choice test to indicate learning with discrimination of the target items.

Based on earlier SL studies and preliminary results from a pilot version of the current study (https://osf.io/bsyz3/), we expected DSL to be a modality-sensitive and stimulus-sensitive ability (as opposed to a unitary one). As such, we expected to find variable performance across modalities and across tasks.

In the categorization task, we predicted higher categorization accuracy in the auditory modality compared to the visual modality due to the increased processing ability of durational information in the auditory modality (e.g., Grondin, 2010). Alternatively, it was also possible to see high accuracy scores in the Visual-Spatial task (but not the Visual-Temporal task, which is predicted to elicit the lowest performance) which would contribute to higher accuracy in the visual modality compared to the auditory modality. Within the auditory modality, we predicted higher categorization accuracy for linguistic stimuli compared to non-linguistic stimuli due to prior exposure to meaningful durational differences in speech input (e.g., Siegelman et al., 2018). Within the visual modality, we predicted higher categorization accuracy for spatial stimuli compared to temporal stimuli as previous work has shown that people are significantly better at learning and processing visual-spatial relationships compared to temporal ones (e.g., Conway & Christiansen, 2009).

Following Frost et al. (2015), we did not predict any significant correlations between individuals' categorization accuracy across modalities and stimuli, supporting the general hypothesis that DSL is a modality- and stimulus-sensitive ability. Note that low but significant correlations across conditions may suggest that there is both a shared component between the modalities and an independent component per modality. Considering that task reliability is critical for interpreting correlational findings, we also investigated the reliability of all tasks using a split-half correlation measure.

In the production task, we predicted that participants would significantly exaggerate the learned categories, rather than reproduce them perfectly. This is because adult learners regularize variable input by using a maximization strategy in tasks that require high cognitive load (e.g., Ferdinand, Thompson, Kirby, & Smith, 2013; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009). We predicted that this exaggeration will occur in one of two possible ways. The first possibility is significant deviation from the category mode: i.e., for the "long" category, participants would produce signals that are significantly longer than those they learned; for the "short" category, participants would produce signals that are significantly shorter than those they learned. The second (and not mutually exclusive) possibility is significant exaggeration of the gap between the "short" and "long" categories, i.e., the mean difference between participants' short and long productions would be greater than the difference between these categories in the original stimuli. For production performances across modalities, we predicted no significant correlations between individuals' production deviation across modalities and stimuli.

We also expected to find a correlation between participants' categorization and production behavior, namely, that categorization accuracy would determine participants' production behavior such that the amount of exaggeration in participants' productions would be significantly and positively correlated with how well they learned and categorized the original stimuli (i.e., better learning of the stimuli would lead to more deviation). This prediction is motivated by literature on rule-learning and categorization (e.g., Kuhl, 1991), which argues that the higher-level abstract representations formed from specific exemplars are often exaggerated with respect to the relevant/distinctive categorical features, i.e., making the prototypical example of category X more X-like. We also considered the possibility that higher categorization accuracy would in fact be associated with *more* precise productions (i.e., better learning would lead to less deviation). This result would support an exemplar-based, associative learning model for SL, which stores individual exemplars and does not readily form abstract categorical representation (i.e., "non-analytical learning", following Smith et al., 2012).

Data availability

Data per individual participant and analysis scripts are accessible via our repository on the Open Science Framework: https://osf.io/vx35h/.

Methods

Design

The experiment was a within-subjects study with four experimental conditions: Auditory-Linguistic, Auditory-Non-Linguistic, Visual-Temporal, and Visual-Spatial. Each experimental condition consisted of three parts: a passive exposure phase (44 trials), a categorization task (55 trials), and a production task (16 trials). In each experimental condition, participants learned to distinguish between two categories (i.e., dangerous and safe aliens) by learning their corresponding signals (i.e., 'Short' or 'Long').

The experiment was semi-randomized without replacement, with the following constraints: *For every participant*, the order of the four experimental conditions was randomized under the constraint that the two auditory conditions cannot occur consecutively; the mapping between signals (short vs. long) and meanings (dangerous vs. safe alien) was randomized across the four experimental conditions under the constraint that the same mapping (e.g., long = safe) cannot occur in more than three experimental conditions; the order of the trials in the exposure phase was randomized under the constraint that the same signal cannot occur twice in a row.

We examined participants' performance on the categorization and production tasks across the different modalities. For the categorization task, we measured categorization accuracy. For the production task, we measured participants' production deviation from the category modes (i.e., how much do they deviate from the category modes in absolute terms), and participants' gap between the short and long categories (i.e., do they exaggerate or minimize the contrast between the categories in relative terms).

Participants

Participants were recruited through Prolific, an online participant database. Participants were paid GBP 7.50 for their participation. All participants were over 18 years old, and had no self-reported uncorrected visual or hearing difficulties. They were all English speakers with different language backgrounds.

A total of 145 people participated, but 7 participants were excluded due to clicking on the same answer more than 20 times in a row in the categorization task (which indicates that they did not understand the task or did not perform it with sufficient intention, see preregistration). A further 19 participants were excluded because their screen resolution was below the minimum of 1024 x 768 px or they did not use a laptop or desktop (with the exception of one participant being kept for Visual-Temporal condition in both tasks because they sufficed the screen resolution criteria only in that condition), leaving a total of 118 participants for analysis (and 119 for the Visual-Temporal condition only).

Materials and stimuli

Two variables were manipulated for the stimuli: signal type and signal length. Each experimental condition had its own signal type (see Fig. 1). Auditory-Linguistic: a syllable [ʔaː] with a given duration; Auditory-Non-Linguistic: a simple tone with a given duration; Visual-Temporal: a red coloration that flashes with a given duration; Visual-Spatial: a flower stem with a given height.

The auditory stimuli were created in PRAAT (Boersma & Weenink, 2019). Specifically, the Auditory-Linguistic stimuli were based on a recording of a glottal stop followed by a long open-front unrounded vowel [ʔaː]. This was split into a consonant and a vowel part by inspection of the spectrogram. The length of the vowel part was then changed by a PRAAT script in such a way that the total stimulus had the desired length. The Auditory-Non-Linguistic stimuli were single tones of 344 Hz. The visual stimuli were clipart images of a mushroom for the Visual-Temporal condition, and a sunflower head with a green rectangle of variable length (i.e., the stem), which was drawn by using the polygon component in PsychoPy.

In the experimental conditions that used duration as the defining feature of the signals, the durations ranged from 313 ms to 811 ms in increments of 10 %. A typical signal from category 1 ('Short') had a duration of 416 ms; a typical signal from category 2 ('Long') had a duration of 609 ms (see Fig. 2). The categories were characterized by means 420 ms and 615 ms, and standard deviations 57 ms and 83 ms (where the standard deviations are different because the durations were spaced exponentially). These distributions were identical to those in the pilot version of the study, ensuring their learnability.

In the Visual-Spatial experimental condition, the unit of measurement was pixels. The heights of the sunflower stems ranged from 78 to 203 px in 10 % increments, which is ¼ * [duration in ms] of the experimental temporal conditions. A typical 'Short' stem was 104 px; a typical 'Long' stem was 152 px. The categories were characterized by means 115 and 154 px and standard deviations 17.5 and 20.8 px, respectively. This range of heights was chosen for a number of reasons. First, there was a proportional correspondence between these heights and the durations in ms. Second, this range allowed participants to exaggerate categories. Finally, the range was suitable for presentation of the images on both high-resolution screens (e.g., 1920 x 1080) and lower-resolution screens (e.g., 1024 x 768). The circle that indicates the

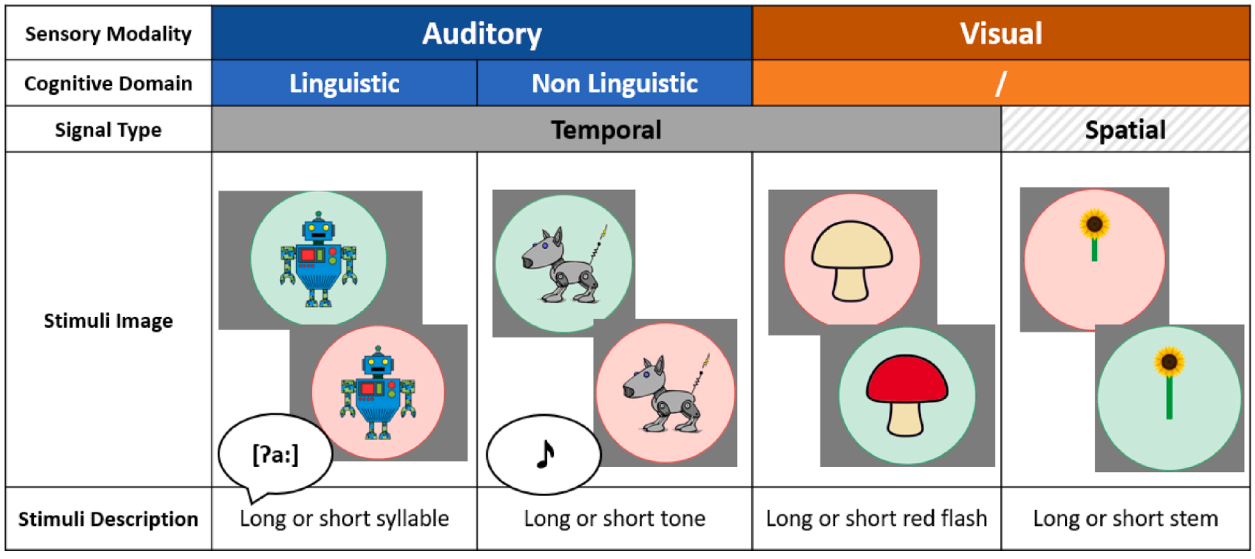

Fig. 1. The four experimental conditions and demonstration examples for their corresponding audio and visual stimuli.

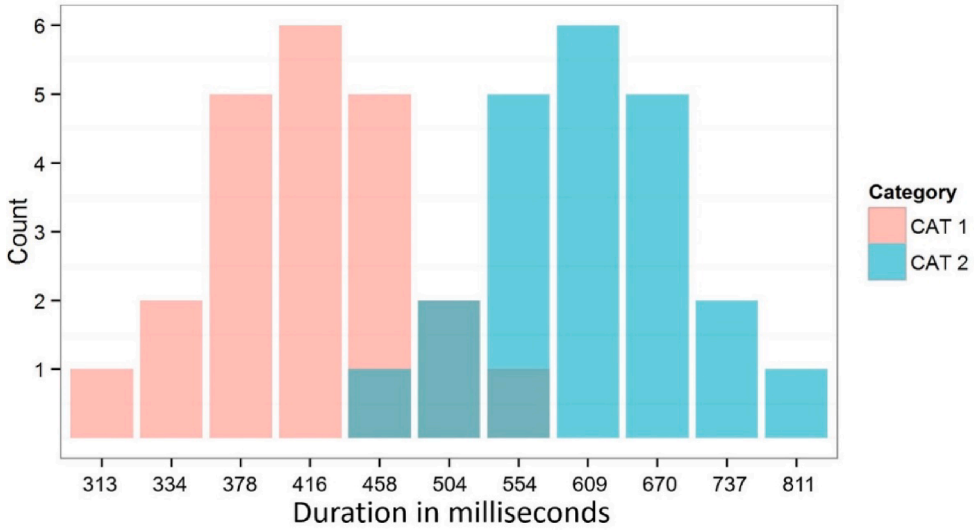

Fig. 2. Distribution of the 'Short' (Category 1) and 'Long' (Category 2) signals in the exposure phase of temporal tasks. The x-axis represents the duration of the signals in the two auditory tasks and in the visual-temporal task. The corresponding signal heights in the Visual-Spatial task are: 78, 86, 95, 104, 115, 126, 139, 152, 168, 185, and 203 px.

category was always 500 px in diameter, and was presented in the middle of the screen (i.e., position 0,0). The flowerhead (size: 120 * 120 px) was always positioned at 125 px above the middle (i.e., position (0,125). The midpoint of the stem always falls at 60–2/[stem height] px.

Procedure

The experiment was created using PsychoPy experimental software (Peirce, 2007), and run through Pavlovia, the repository and launch platform for PsychoPy experiments. Participants were required to use a personal computer (instead of a phone or tablet).

After signing the consent form, participants were asked some background questions: their age, which language(s) they speak, whether they have a learning disability (yes or no), whether they have been diagnosed with AD(H)D (yes or no), what type of device they are using, and whether the device uses a mouse, touchpad, or touch screen. The screen resolution was recorded to make sure that participants were using a screen that was compatible with the task.

In the exposure phase, participants were exposed to two alien categories that send out signals of variable duration or height. For the two auditory conditions, participants heard the stimuli through the speakers of their device or through headphones. For the visual conditions, participants saw the stimuli on their computer screens. During a given trial, the signal was presented simultaneously with an image that corresponds to the category it belonged to. Specifically, each signal was accompanied by a green or a red circle that indicated whether the signal came from a 'safe' alien (green circle) or a 'threat' (red circle) (see Fig. 3A). The exposure phase was defined by an overlapping one-dimensional Gaussian probability-density function and was represented by 'Short' (Category 1) or 'Long' (Category 2) signals (see Fig. 2). There were 11 unique signals in this signal space. The most extreme signals were only presented once, while the typical signals were presented 6 times, resulting in 44 exposure trials per participant.

In the temporal conditions, an exposure trial showed the images of the alien and the circle for 2000 ms, and the signal always started 500 ms after the start of the images. After each stimulus presentation, a blank screen was shown for 400 ms, after which the next trial started (see also Fig. 3A). In the visual-spatial condition, an exposure trial showed both the category and the stimulus for 500 ms. This duration was chosen because duration was fixed in this experimental condition, and 500 ms was around the midpoint of the continuum for the experimental temporal conditions. After each presentation, a blank screen was shown for 1500 ms, so that each trial lasted 2000 ms in total, after which the next trial started. As the trial ended with a blank screen, the additional blank screen of 400 ms (as in the temporal conditions) was not added in this experimental condition.

Following the exposure phase, participants were tested with a categorization and a production task. Responses were submitted using a mouse or a touchpad: In the categorization task, participants selected their answer on the screen by clicking on it. In the production task, participants clicked their mouse or their touchpad to produce the signals.

The categorization phase was the same in all four experimental conditions: signals were presented and categorized one by one. Presentation was exactly as it was in the exposure phase, except that the red or green circles were replaced with a white one. After the presentation of each signal, participants categorized the signal using a 6-point Likert scale (see Fig. 3B). Before selecting their answer, participants could choose to replay the signal once. Clicking the Likert scale ended the trial. Participants did not receive feedback on the accuracy of their answers. Each of the 11 signals was categorized 5 times, resulting in 55 categorization trials per participant.

In the production task, participants had to produce a typical signal for the two categories. In each trial, the target category was presented (i.e., the base image surrounded by a red or green circle), and participants could produce the corresponding signal by pressing and holding a button (see Fig. 3C). The signal would be presented for the entire duration of the button press, and up to 2500 ms max. In the duration conditions, participants can only press this button and hold once. Once released, participants could click the 'next item' button. In the spatial task, participants could release and then re-press the button to further increase the height of the sunflower stem (i.e., they cannot make it smaller or redo the production). This change was included because pilot participants indicated that they often released the button too early, and therefore were unable to submit signals that reflected their representations of the categories.

At the end of the experiment, each participant was asked which experimental condition they found the most difficult, and which they found the easiest (required question). We also asked them to share their learning strategy, should they have one (optional). Finally, they could add any comments and questions if they have any (optional).

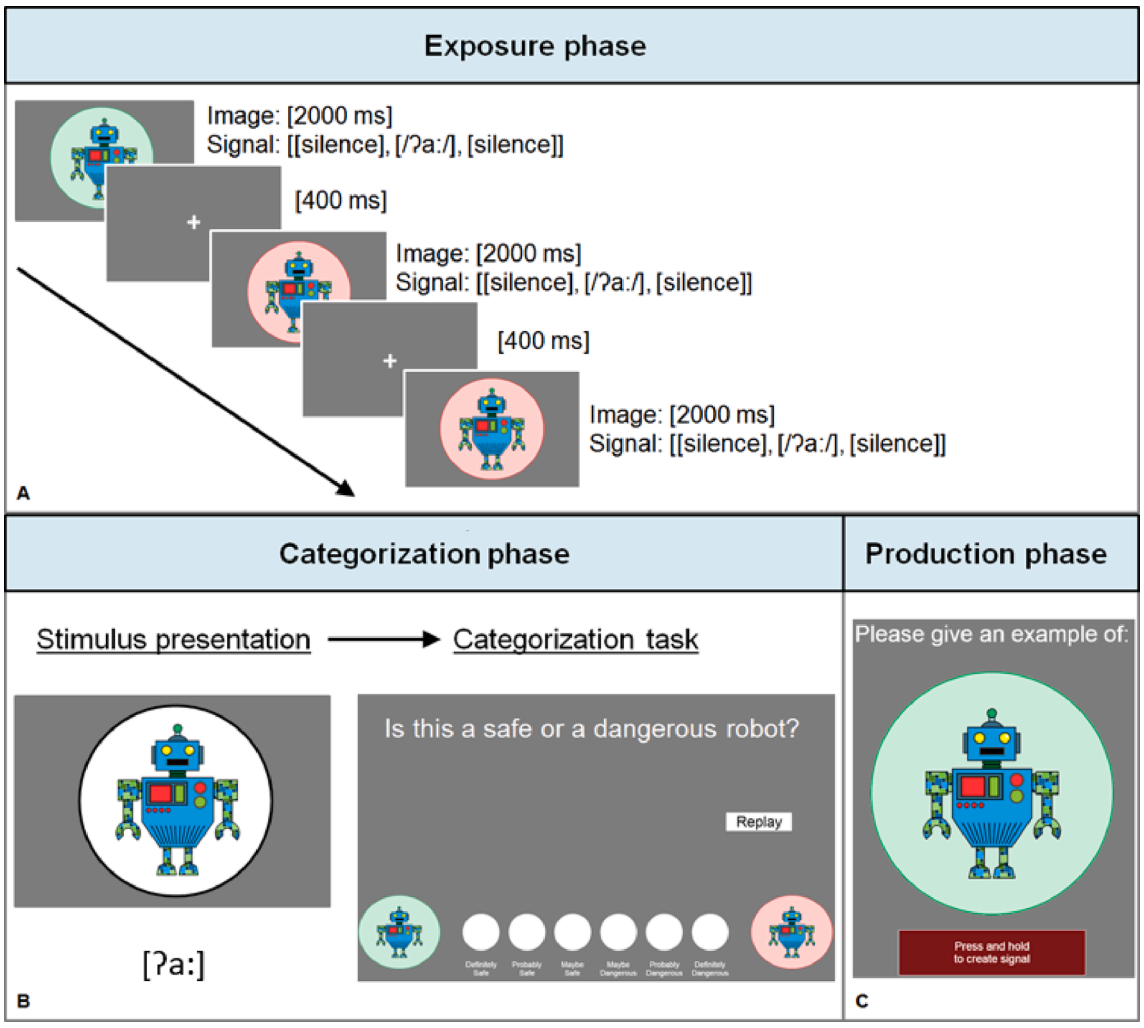

Fig. 3. The three phases of the experiment for the Auditory-Linguistic condition. Exposure phase (panel A), Categorization phase (panel B) and Production phase (panel C). A: In the Exposure phase, a trial is 2000 ms during which the image is shown. The signal is always presented 500 ms after the image onset, and ends before the end of the trial. B: In the categorization phase, the stimulus is always presented before the categorization task. If participants press the replay button, the image is presented in the area above the 6-point Likert scale of the Categorization task screen. C: In the production phase, the red rectangular button must be pressed to create a signal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Analyses

Measuring categorization accuracy

Categorization accuracy was measured in the following way: First, we calculated the probability of a given signal being assigned to the 'Short' category during the training phase (see Fig. 2). Specifically, in the experimental conditions in which duration was manipulated, signals with a duration of 313 to 416 ms were always assigned to 'Short' (i.e., 100 % of the time); signals with a duration of 458 ms were assigned to 'Short' 83.33 % of the time (5/6 presentations); signals with the duration of 504 ms were assigned to 'Short' 50 % of the time; signals with the duration of 554 ms were assigned to 'Short' 16.67 % of the time (1/6 presentations); and signals with the duration of 609 to 812 ms were never assigned to 'Short' (0 % of the time). In the Visual-Spatial task, the corresponding transformation was as follows: 78 to 104 px (100 %), 115 px (83.33 %), 126 px (50 %), 139 px (16.67 %), and 152 to 203 px (0 %). Then, for each trial in the categorization testing phase, we calculated participants' accuracy based on their chosen category, their confidence, and the probability of the given signal being assigned to the selected category, using the following symmetrical scoring method (Table 1):

We used this scoring method because it captures the likelihood of a specific response being given based on the probability of the signal's category during training[2] (see Appendix C for more details).

Measuring production performance

Production performance was measured in two ways. The first measure, production deviation, was calculated as the deviation of the produced signal from the category modes (i.e., the 'peaks' of the distributions). In the experimental conditions in which the duration of the signal was manipulated, production deviation was measured in milliseconds. The peaks were at 416 ms for the 'Short' category and 609 ms for the 'Long' category. In the Visual-Spatial task, production

[2] We also used transformed Brier scores (a standard method for measuring the accuracy of probabilistic predictions, see Appendix C) to calculate participants' accuracy. In the pilot study, the Brier scoring method yielded similar results to our own scoring method, but the former was dispreferred given its less subtle and more conservative treatment of uncertainty. For example, when the signal itself was ambiguous (i.e., categorized at short 50% of the time during training), the transformed Brier scores assigned participants with a maximal accuracy of 0.49, even when they chose the so-called "correct" uncertain response ("Maybe short" or "Maybe long"). As such, we chose to use the scoring method depicted in Table 1 for the current study, which captures the intuitive accuracy scheme.

Table 1
Illustration of the scoring method used for the categorization task.

	Probability of target signal being assigned to 'Dangerous' category				
	0%	*16.67%*	*50%*	*83.33%*	*100%*
Participants' response	***Assigned accuracy***				
Definitely 'Dangerous'	*0*	*0*	*0.2*	*1*	*1*
Probably 'Dangerous'	*0.2*	*0.2*	*0.6*	*1*	*0.8*
Maybe 'Dangerous'	*0.4*	*0.4*	*1*	*0.6*	*0.6*
Maybe 'Safe'	*0.6*	*0.6*	*1*	*0.4*	*0.4*
Probably 'Safe'	*0.8*	*1*	*0.6*	*0.2*	*0.2*
Definitely 'Safe'	*1*	*1*	*0.2*	*0*	*0*

deviation was measured in px. The peaks of the distributions were at 104 px for the 'Short' category, and 152 px for the 'Long' category. A positive production deviation indicates that participants were exaggerating the signals (i.e., producing an even shorter signal for the 'Short' category compared to its mode, or producing an even longer signal for the 'Long' category compared to its mode).

The second measure, gap difference, was how much participants exaggerate the difference between the 'Short' and 'Long' categories with their reproductions. In the training data, the gap between the category modes is 193 ms for the temporal conditions or 48 px for the spatial conditions. A positive gap difference indicates that the reproduced gap was exaggerated from the modal differences between the two categories.

For both measures, responses in the spatial condition were multiplied by 4 to match the scale of the temporal conditions.

Linear mixed-effects models

We examined participants' performance on the categorization and production tasks across the different experimental conditions. We also tested the relationship between participants' performance on these two tasks across conditions. These analyses were done using linear mixed-effects (LME) regression models generated by the lme4 and lmetest packages in R (Bates, Mächler, Bolker, & Walker, 2014; R Core Team, 2021; Zeileis & Hothorn, 2002). All regression models included a fixed effect for Experimental Condition, which is a 4-level categorical variable with user-defined contrasts, coded to make the following comparisons: Auditory-Linguistic vs. Auditory-Non-Linguistic condition; Visual-Temporal vs. Visual-Spatial condition; Visual-Temporal vs. the two Auditory conditions. The last contrast captures the difference between the auditory and the visual modalities with respect to temporal stimuli.

As for random effects, we aimed to use the maximal random effect structure justified by the design and our hypotheses, i.e., random intercepts for participants and signal duration, as well as by-participant and by-signal random slopes with respect to the main effect of condition. In case the model did not converge with this maximal structure, we removed the random effects with the lowest variability, one at a time (Barr, Levy, Scheepers, & Tily, 2013). We used the p-values generated by the lmetest package in R, where we interpreted $p < 0.05$ as indicating that the specified fixed effect estimate is significant (Zeileis & Hothorn, 2002).

Correlations

Prior to the correlational analyses we documented the split-half reliability of each task. A permutation-based split-half approach with 5000 random splits was adopted to estimate the internal consistency of each experimental condition. The estimations were done using the splithalf package in R (Parsons, 2021; R Core Team, 2021).

We then used Pearson's product-moment correlation to examine the relation between individuals' scores in the different experimental conditions. We removed bivariate outliers for each correlation test using the Mahalanobis-distance method with a breakdown point of 0.25 (outliers_mcd function in the Routliers library, Leys, Klein, Dominicy, & Ley, 2018). The number of outliers removed for each correlation test can be seen in Table B.1 in Appendix B. We used the p-values generated by the Hmisc package in R (Harrell & Dupont, 2023) and given that we looked at multiple correlations, the p-values were adjusted according to Bonferroni correction.

Results

Results are reported for the statistical analyses of the categorization task (accuracy), the production task (deviation and gap difference), and the relation between the categorization and production measures. As a guide to what is a substantial set of analyses, we expected significant differences and no meaningful correlations across modalities and stimulus types for all three task measures. Additionally, we predicted a significant effect of categorization accuracy on the production task measures.

Categorization task

Overall, the 118 participants achieved 80.84 % mean accuracy for the categorization task. For each experimental condition, the mean accuracy score was 82.27 % (SD = 0.27) in Auditory-Linguistic, 81.38 % (SD = 0.27) in Auditory-Non-Linguistic, 76.14 % (SD = 0.30) in Visual-Temporal, and 83.56 % (SD = 0.27) in Visual-Spatial (see Fig. 4).

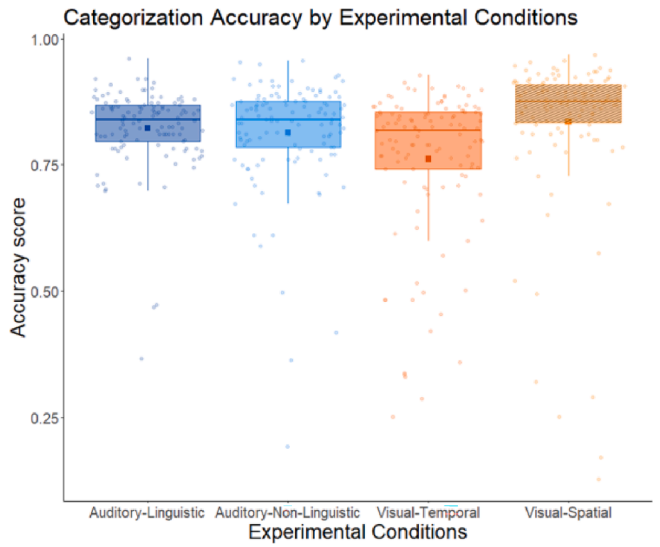

Fig. 4. Categorization accuracy by experimental condition. The box captures 50% of data in each condition with the upper and the lower hinges corresponding to 25th and 75th percentiles. The whiskers capture the data within 1.5 interquartile range from the hinges. The line in each box is the median of the accuracy scores and the square indicates the mean for each experimental condition. Dots represent individual participants.

A LME model was used to predict categorization accuracy with experimental condition as a fixed factor.[3] In the final model, random intercepts for participants and signal durations were included. The model syntax and the full results table are in the Appendix A (Model Eq. (A.1)). More importantly, we found no significant difference between the linguistic and non-linguistic auditory conditions. Comparing the two visual conditions, accuracy in the Visual-Temporal condition was significantly lower than that in the Spatial condition ($\beta = -0.07$, $t = -3.83$, $p < 0.001$). Finally, comparing modalities, accuracy in the Visual-Temporal condition was significantly lower than that of the two auditory conditions ($\beta = -0.06$, $t = -4.18$, $p < 0.001$).

Since the signal durations were consistent throughout the three temporal conditions, an additional ANOVA test was conducted to check if there was an effect of order on the categorization accuracy. No significant improvement of performance was found ($F(2,352) = 0.13$, $p = 0.88$) across the three conditions. Tukey's HSD Test for multiple comparisons showed that there was no significant difference between the 1st (mean = 0.79) and the 2nd (mean = 0.80) conditions ($p = 0.94$, 95% C.I. = [−0.03, 0.04]), or between the 1st and the 3rd (mean = 0.80) conditions ($p = 0.88$, 95% C.I. = [−0.03, 0.04]).

Prior to the correlational analyses we evaluated the split-half reliability for all conditions of the categorization task (see Table 2). Spearman-Brown corrected split-half correlation coefficients were above 0.80 for all conditions of the categorization task, indicating good task reliability so that it is meaningful to investigate the correlations between conditions.

As seen in Fig. 5, the accuracy scores for conditions within the same modality significantly correlated with each other. Across the two modalities, categorization accuracy of auditory conditions significantly correlated with that of the Visual-Temporal condition.

[3] Since there is a sizeable negative skew in the categorization data, particularly in the Visual-Temporal condition, we also conducted a Friedman rank sum test to compare the medians across the four experimental conditions (Auditory-Linguistic: 0.84, Auditory-Non-Linguistic: 0.84, Visual-Spatial: 0.88, Visual-Temporal: 0.82). One participant lacked data for three out of four experimental conditions and was therefore removed to meet the complete block design assumption for the test. The test showed significant difference between the medians across all conditions, $F_r = 60.209$, df = 3, $p < .001$. For a follow-up pairwise comparison, two Wilcoxon's signed ranks tests (Auditory-Linguistic vs. Visual-Temporal: $Z = 0.85$, $p < .001$; Auditory-Non-Linguistic vs. Visual-Temporal: $Z = 0.79$, $p < .001$) and another Friedman's test (contrasts between Auditory-Linguistic, Auditory-Non-Linguistic, and Visual-Temporal: $F_r = 12.197$, df = 2, $p < .05$) comparing the Visual-Temporal condition with the two auditory conditions directly showed significant difference between the experimental conditions.

Table 2
Split-half reliability results for the categorization task in each of the four experimental conditions.

	Auditory-linguistic	Auditory-non-linguistic	Visual-temporal	Visual spatial
Participants (N)	118	118	119	118
Split-half coefficient	r = 0.70 95CI (0.62–0.77)	r = 0.84 95CI (0.79–0.88)	r = 0.89 95CI (0.86–0.92)	r = 0.92 95CI (0.90–0.94)
Spearman-Brown correction	r = 0.82 95CI (0.76–0.87)	r = 0.91 95CI (0.88–0.93)	r = 0.94 95CI (0.92–0.96)	r = 0.96 95CI (0.95–0.97)

Production task

Fig. 6 presents the production responses by signal categories ('Short' or 'Long') and experimental conditions. The figure shows that the produced auditory signals are longer than they actually were, whereas that was not the case for visual signals. To examine this trend, we looked at two production measures: (1) Production deviation (from category mode), which was the distance between the average response length (for short or long signals) and the corresponding mode line (416 for short and 609 for long); and (2) Gap difference, which was calculated by first obtaining the gap of average response length between the long and short signal categories and then compare that to the gap between the modal signals (i.e., 193). The average production deviation was 458.25 across the four experimental conditions (Auditory-Linguistic: 983.66, SD = 610.19; Auditory-Non-Linguistic: 868.22, SD = 506.45; Visual-Temporal: 7.45, SD = 406.58; Visual-Spatial: −26.33, SD = 296.06). The mean gap difference across experimental conditions was 219.84, with a difference score of 242.4 (SD = 458.51) in Auditory-Linguistic, 225.45 (SD = 370.02) in Auditory-Non-Linguistic, 209.04 (SD =

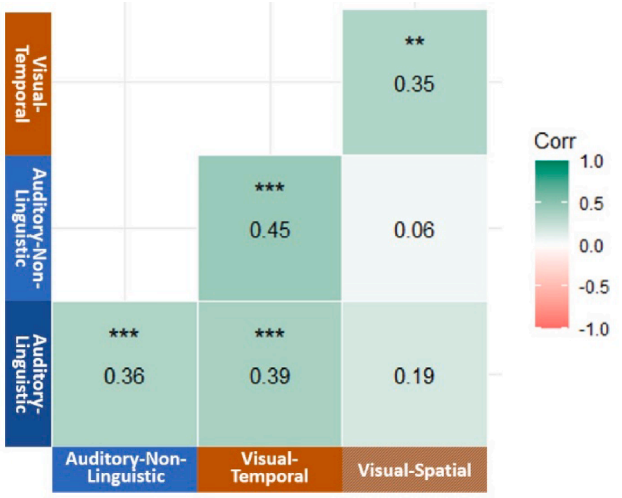

Fig. 5. Correlation matrix for categorization accuracy across experimental conditions. The r-values in the plot reflect results after bivariate outlier removal. All p-values were corrected for multiple comparisons using Bonferroni correction. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$. Scatter plots and r-values without outlier removal can be found in Appendix B, Figure B.1.

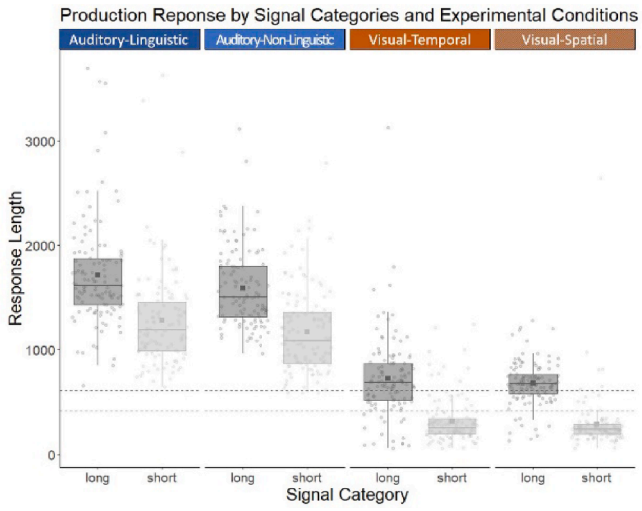

Fig. 6. Production response by signal categories and experimental conditions. Data of the Visual-Spatial condition was rescaled (multiplied by four) to allow for comparison between the temporal and spatial tasks. The box captures 50% of data in each condition with the upper and the lower hinges corresponding to 25th and 75th percentiles. Whiskers capture the data within 1.5 inter-quartile range from the hinges. The line in each box is the median of response means among all participants. The square inside the box indicates the mean for each category and condition. The dotted line indicates the category mode in dark grey for the 'Long' category and in light grey for the 'Short' category. Dots represent individual participants.

474.58) in Visual-Temporal, and 202.47 (SD = 378.38) in Visual-Spatial conditions.

A LME model was used to predict production deviation from mode (Model Eq. (A.2) in Appendix A). The final model had experimental conditions, category length, and their interaction as fixed factors. A full random effect structure for participants was also included. The model showed a significant intercept for deviation from mode ($\beta = 142.11$, $t = 23.15$, $p < 0.001$). There was a significant difference between auditory conditions ($\beta = 31.66$, $t = 3.05$, $p < 0.01$) but not visual conditions. A significant difference was also shown across modality conditions: Visual-Temporal production deviation was significantly smaller than the auditory production deviations ($\beta = -233.01$, $t = -24.06$, $p < 0.001$). In addition, there was a significant difference in the production deviation based on category length $\beta = -54.44$, $t = -7.92$, $p < 0.001$). The deviation was significantly larger for the long categories than for the short ones.

We used another LME model to predict the effect of experimental conditions on gap difference (Model Eq. (A.3) in Appendix A). The model included random intercepts for participants but no random effect of category's true length or by-participant slopes. Results showed a significant intercept but no significant difference between stimuli types or modality ($\beta = 54.13$, $t = 7.80$, $p < 0.001$).

We investigated the production results per conditions using a split-half reliability correlation measure (Table 3). The Spearman-Brown corrections were all 0.75 or higher.

We also investigated the correlations between deviation from mode across experimental conditions using Pearson's test with the multivariate outliers removed accordingly (Fig. 7). After bivariate outlier removal, the scores of deviation from mode for the two auditory conditions and the two visual conditions were significantly correlated. Thus, the production performance showed correlations exclusively within each modality condition. Although after Bonferroni correction, only the Auditory-Linguistic and Auditory-Non-Linguistic correlation stayed significant ($p < 0.001$).

Fig. 8 displays the correlation results for the gap difference measure across experimental conditions. Unlike the production deviation from

Table 3
Split-half reliability results for the production task in each of the four experimental conditions. Panel a. reports results for the measure of **production deviation from mode**; panel b. for the measure of **gap difference**.

a.	Auditory-linguistic	Auditory-non-linguistic	Visual-temporal	Visual-spatial
Participants (N)	118	118	119	118
Split-half coefficient	r = 0.86 95CI (0.81–0.90)	r = 0.86 95CI (0.81–0.90)	r = 0.70 95CI (0.53–0.81)	r = 0.61 95CI (0.42–0.73)
Spearman-Brown correction	r = 0.93 95CI (0.90–0.95)	r = 0.92 95CI (0.90–0.95)	r = 0.82 95CI (0.69–0.89)	r = 0.75 95CI (0.58–0.84)
b.	**Auditory-Linguistic**	**Auditory-Non-Linguistic**	**Visual-Temporal**	**Visual-Spatial**
Participants (N)	118	118	119	118
Splithalf coefficient	r = 0.70 95CI (0.63–0.76)	r = 0.69 95CI (0.59–0.78)	r = 0.87 95CI (0.73–0.94)	r = 0.92 95CI (0.90–0.94)
Spearman-Brown correction	r = 0.82 95CI (0.77–0.87)	r = 0.81 95CI (0.74–0.88)	r = 0.93 95CI (0.84–0.97)	r = 0.96 95CI (0.95–0.97)

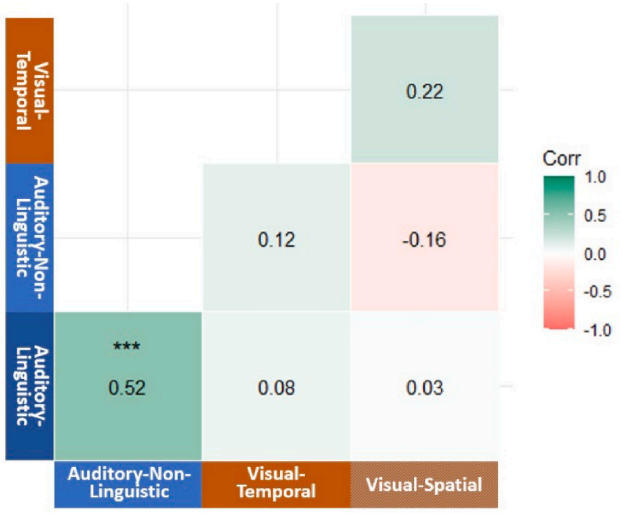

Fig. 7. Correlation matrix for deviation from mode measure across experimental conditions. The r-values in the plot are results after bivariate outlier removal. All p-values were corrected for multiple comparisons using Bonferroni correction. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$. Scatter plots and r-values without outlier removal can be found in Appendix B, Figure B.2.

mode, the gap difference showed significant correlations within both modality conditions (Auditory: $r = 0.47$, $t(1\ 0\ 1) = 5.33$, $p < 0.001$; Visual: $r = 0.51$, $t(1\ 1\ 6) = 3.50$, $p < 0.001$). There were significant correlations between the temporal conditions as well (Auditory-Linguistic and Visual-Temporal: $r = 0.37$, $t(98) = 3.98$, $p < 0.001$; Auditory-Non-Linguistic and Visual-Temporal: $r = 0.50$, $t(97) = 5.68$, $p < 0.001$). Overall, the pattern of correlation results for category difference across experimental conditions resembles that for the categorization accuracy.

Between production deviation and categorization accuracy

Two LME models were used to predict production performances with fixed effects of categorization accuracy and experimental conditions (Model Eq. (A.4) & Eq. (A.5) in Appendix A). The model for production deviation of mode included random intercepts for participants and true

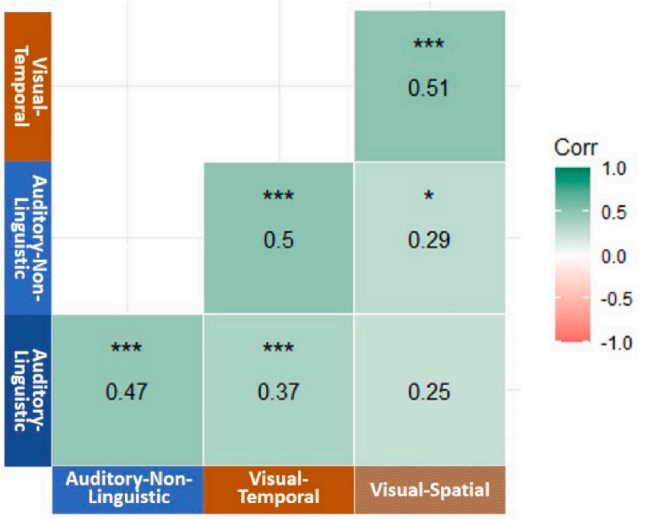

Fig. 8. Correlation matrix for gap difference measure across experimental conditions. The r-values in the plot are results after bivariate outlier removal. All p-values were corrected for multiple comparisons using Bonferroni correction. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$. Scatter plots and original r-values can be seen in Appendix B, Figure B.3.

category length, as well as a by-participant random slope for the effect of modality. Since the model with maximal random effect structure failed to converge (singular fit), we had to remove the by-length random slope for the effect of modality. Unlike the prediction of better learning leading to more or less deviation in reproductions, no significant effect of categorization accuracy was found ($\beta = 4.83$, $t = 0.14$, $p = 0.89$). The model for production category difference included only random intercepts for participants due to convergence failure of the maximal structure. The results showed a significant effect of categorization accuracy as was predicted ($\beta = 259.37$, $t = 6.70$, $p < 0.001$).

Discussion

The goal of the current preregistered study was to investigate modality and stimulus differences in DSL by comparing the learning of categories from a continuous signal range (based on the frequency distribution of short and long signals) in the auditory and visual modalities, as well as across different types of stimuli within a modality. To test whether participants show enhanced performance for any particular type of stimuli, we compared learning with linguistic stimuli (syllable) and non-linguistic stimuli (tone) for the auditory conditions, and with temporal (flash) and spatial (stem shape) stimuli for the visual conditions. Unlike previous SL studies that mostly measure learning with categorization accuracy, our study also included an extra production task to test DSL from an active learning perspective. Specifically, we analyzed the production data using two different measures: production deviation from the category mode, and the difference between the gap of the produced categories and the modal gap. The within-subject design of the experiment allowed us to investigate whether an individual's performance on a task in one modality is predictive of their performance in other modalities (as in Siegelman & Frost, 2015). Moreover, we looked at the relations between different learning measures and asked whether categorization accuracy predicts production behavior. A visualization of the summary of results can be seen in Fig. 9.

Summary of results

While participants showed successful learning of categories in all modality and stimulus conditions, a direct comparison of DSL across conditions revealed some important differences. In line with our predictions based on previous findings on CSL (Conway & Christiansen, 2009; Grondin, 2010), durational information was better learned in the auditory modality, while spatial information was better learned in the visual modality. On the other hand, we did not find the predicted linguistic advantage within the two auditory conditions.

Our design also included a production task. Compared to the original signals from the training distribution, participants in all conditions produced signals that were significantly exaggerated, namely, further from the category modes and with greater differences between categories. Notably, participants who learned better (as measured during the categorization task) also produced a bigger gap between the short and long categories. This suggests that the production measure of gap difference might be a better predictor of learning when compared to the measure of deviation from mode.

On the individual level, we found significant correlations across both modalities and stimuli in the categorization task and in one of our production measures (i.e., gap difference). Participants' learning and reproduction were positively correlated in conditions that shared a modality or when the stimuli were all temporal in nature which aligns with previous correlational findings (Siegelman et al., 2018; Growns

<table>
<tr><th></th><th>Categorization task</th><th colspan="2">Production task</th><th colspan="2">Production ~ Categorization</th></tr>
<tr><th></th><th>Categorization Accuracy (CatAcc)</th><th>Deviation from mode</th><th>Gap difference</th><th>Deviation from mode ~ CatAcc</th><th>Gap difference ~ CatAcc</th></tr>
<tr><td rowspan="3">LME</td><td>AL ≠ ANL</td><td rowspan="3">Significant exaggeration</td><td rowspan="3">Significant exaggeration</td><td rowspan="3">No effect</td><td rowspan="3">Positive significant effect</td></tr>
<tr><td>VS > VT</td></tr>
<tr><td>A > V</td></tr>
<tr><td rowspan="2">Correlation</td><td>Significant correlation between AL/ANL & VT</td><td>No significant correlation across modalities</td><td>Significant correlation between AL/ANL & VT</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Significant correlation between AL & ANL; VS & VT</td><td>Significant correlation between AL & ANL</td><td>Significant correlation between AL & ANL; VS & VT</td></tr>
</table>

Fig. 9. Summary of results. The color of cells indicates whether a prediction is met (green) or not (red). Four experimental conditions are abbreviated as AL (Auditory-Linguistic), ANL (Auditory-Non-Linguistic), VT (Visual-Temporal), and VS (Visual-Spatial). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

et al., 2020). The results of production deviation from mode were slightly different, with a significant correlation only between the auditory conditions. No significant correlation was found across sensory modalities even when they both used temporal stimuli. The two visual conditions did not significantly correlate, which may be due to the lower reliability of the production task in the Visual-Spatial condition.

Overall, the observed pattern of correlations suggests that DSL is characterized by a certain degree of modality- and stimulus-sensitivity. Specifically, conditions that matched in either modality or stimulus type showed higher correlations than those that did not. However, our results do not support a strong modality-specific or stimulus-specific scenario, as such accounts would predict no correlations across modalities or tasks with different stimulus types (cf. Frost et al., 2015). Notably, the effects of modality and stimulus type were not additive; a match for both modality and stimulus type did not further increase the correlations found between conditions. In the following discussion, we will link our findings to current accounts of SL, and discuss potential limitations of our experiment.

Modality- and Stimulus-Specificity in DSL

A theoretical account

Previous CSL studies found that people do not show a uniform sensitivity to statistical regularities across different input modalities (e.g., Conway & Christiansen, 2005; Emberson et al., 2011) and this independence was shown at both group and individual levels (Siegelman & Frost, 2015). Nevertheless, correlations between auditory and visual CSL performances were present in more recent studies (e.g., Siegelman et al., 2018). To account for both types of findings, Frost et al. (2015) proposed that individual SL performance results from constraints determined by specific input properties and domain-general computational mechanisms. Specifically, the process of encoding an internal representation of a given stimulus would go through computations instantiated in the designated cortical areas (e.g., auditory and visual cortex) and thus naturally takes the input modality into account. Since each sensory cortex is sensitive to different types of information, the encoding process will be affected accordingly. For example, the auditory cortex is more sensitive to temporal vs. spatial information, whereas the visual cortex is more sensitive to spatial vs. temporal information (Chen & Vroomen, 2013; Conway & Christiansen, 2009).

Our findings are consistent with Frost et al. (2015)'s theoretical account as we observed moderate correlations that suggest some level of generality in our DSL data, while also revealing constraints imposed by input modality and stimulus type. The auditory stimuli were better categorized than the Visual-Temporal stimuli, likely due to the different sensitivity between auditory and visual cortex to temporal information. Regarding the effect of stimulus type, there was indeed a higher categorization accuracy in the Visual-Spatial than the Visual-Temporal condition. These effects of modality and stimulus align with those previously found for CSL (Conway & Christiansen, 2005; 2009; Emberson et al., 2011). The alignment suggests that, despite the potential differences in their underlying memory process (Thiessen et al., 2013), DSL and CSL share similar constraints of modality and stimulus type. It may imply that both types of SL are supported by a unified mechanism. However, since both DSL and CSL commence with an encoding phase where modality and stimulus constraints may already be at play, the communal effect of constraints on learning outcomes could be mediated through the shared encoding component of the two SL processes.

On another note, recent literature suggests that simple learning mechanisms such as Hebbian learning, typically linked to associative learning (Arndt, 2012), may also underlie SL (e.g., Goujon, Didierjean, & Thorpe, 2015; Schapiro & Turk-Browne, 2015; Endress & Johnson, 2021). In our task, the observed learning performance could in principle be explained by the learning of simple associations between the unambiguous signals and the labels of "safe" and "dangerous". Although the positive prediction of significant exaggeration in gap difference by categorization accuracy may suggest formation of categorical representations rather than learning of individual exemplars (as hypothesized based on Kuhl, 1991 in The Current Study), it remains crucial for future research to further elucidate the mechanism(s) underlying the learning of novel categories. One concrete avenue would be to focus on the learning of categories with more overlap, increasing the ambiguity of signal-to-label mappings.

Effect of previous experiences

It can be noted that the higher categorization in Visual-Spatial than Visual-Temporal stimuli may also be attributed to a familiarity effect. In their study targeting visual CSL, Perfors and Kidd (2022) demonstrated that the outcome of SL depends on individuals' perceptual fluency, their ability to quickly and efficiently encode the input. Since both modality and stimulus type can affect perceptual fluency, the particular input which is easier for participants to encode should be learned better. In our study, the superior learning observed for our Visual-Spatial stimuli may result from its natural occurrence in real-life settings, which makes it more familiar for participants compared to the relatively unnatural Visual-Temporal input.

Although a familiarity effect was suggested considering the difference between our experimental design and the real-life language setting, our participants did not show an advantage in learning linguistic over non-linguistic input as was found in previous CSL works (see Siegelman et al., 2018). Originally, we predicted that due to participants' prior exposure to meaningful durational differences in speech input (and particularly with different vowel lengths as being different phonemes), they would show better categorization and production of linguistic stimuli. However, participants in our experiment performed equally well on both linguistic and non-linguistic tasks. The lack of a linguistic advantage could potentially be attributed to the difference between the stimuli used in DSL and CSL studies: our DSL linguistic condition included only one syllable ([ʔaː]), while classic CSL tasks usually include a large set of different syllables (e.g., typically around 10–15 syllables). Thus, participants' prior linguistic knowledge may have a greater impact on CSL tasks simply because there is more linguistic material to be processed (see Siegelman et al., 2018 for discussion). Alternatively, considering participants' overall high performance, the categorization task could be too easy for the participants to accommodate the facilitation of previous experiences.

In the current study, we used durational stimuli for the temporal conditions. In the Auditory-Linguistic stimuli, this meant that the two categories were differentiated by the presence of a long or a short vowel. In some languages, vowel duration is already an important categorical cue that differentiates between phonemes and thus between meanings (e.g., "sika" and "siika" in Finnish, Iivonen & Harnud, 2005). Given that our participants came from diverse linguistic backgrounds, it is possible that there was a non-uniform influence of prior knowledge in the categorization and production of the linguistic stimuli, but not for other stimuli. To further investigate the impact of prior (linguistic) knowledge on the learning of durational categories, future work could directly compare native speakers of languages that either do or do not include durational contrasts in vowels (e.g., Dutch vs. Hebrew, respectively). If prior language experience with vowel duration leads to enhanced performance in linguistic tasks that rely on such a cue for differentiating between categories, we would expect better categorization and production accuracy in speakers of languages such as Finnish or Dutch.

Measuring DSL with production accuracy

As the first study applying a production task to DSL, we analyzed the production data using two different measures that potentially tap into different aspects of production accuracy, i.e., the deviation from mode and the gap difference. However, the results were inconsistent across these two measures which makes it harder to draw clear conclusions with respect to modality and stimuli differences. Notably, unlike the pattern obtained from the categorization task and from our measure of

production gap difference, the measure of production deviation from mode did not show stimulus-specificity (i.e., there *were no* significant correlations across the temporal stimuli for auditory and visual conditions). Participants also had more deviation from mode in the auditory conditions than in the visual conditions, whereas no such difference was found in the measure of gap difference. This inconsistency may stem from differences in task design between modality conditions. While participants were allowed to stop and start the increase of the visual image in the visual tasks, they could only stop once in the auditory tasks. Although this design was refined from the pilot study to accommodate participants' feedback on the visual task (section Procedure), it inadvertently encouraged "over-exaggeration" for the auditory stimulus. Participants could not stop short and continue to increase as the signal approaches the desired audio target as for the visual stimulus. On the other hand, pressing the mouse for reproducing signal durations was an artificial setup, in particular for the Auditory-Linguistic stimuli as it did not correspond to real-life language production. Moving forward, investigations could explore more naturalistic tasks, such as actual syllable production, while ensuring an unbiased production procedure for both auditory and visual conditions.

In addition, production deviation from mode did not correlate with categorization accuracy, further indicating its limited predictive ability of participants' learning. That is, it was not the case that better learning of the categories led to more exaggeration in production, unlike what we found for the measure of gap difference. Our original prediction was based on work by Raviv and Arnon (2018) and Johnson, Siegelman, and Arnon (2020), which revealed that learners who show increased sensitivity to regularities in the input also show more extreme biases for creating structure when reproducing signals they learned during training. Future work should explore such cross-modal correlations for CSL, since a production task may also provide valuable insights into understanding the possible difference in CSL performance caused by task types.

Bearing in mind that SL abilities vary among individuals, it is important to ensure that our measures are reliable and therefore valid for further inferential analysis (Erickson, Kaschak, Thiessen, & Stutts Berry, 2016; Siegelman et al., 2017; Siegelman & Frost, 2015). Although the production task was newly introduced, our split-half results are promising with regard to the reliability of the tasks. Future works should also evaluate test–retest reliability for both old and new measures of cognitive and behavioral performances so that the true effect of the interventions can be better detected and revealed.

Conclusion

This study examined learners' distributional learning using a within-subjects design, comparing performance across different sensory modalities (auditory vs. visual), signal types (temporal vs. spatial), and cognitive domains (linguistic vs. non-linguistic), filling an important gap in the SL literature. We examined participants' learning using two different sources of information: categorization behavior (indicating the degree to which participants have learned the stimuli they were exposed to), and production behavior (indicating which categories were formed by the participants, and how each category is prototypically represented). Our findings revealed that DSL is influenced by modality and stimulus conditions, aligning with prior reports on CSL. Additionally, DSL performance tended to correlate across conditions, suggesting some level of generality in the mechanism. These findings contribute to a more comprehensive understanding of SL by filling in knowledge gaps for distributional learning behaviors. Furthermore, we encourage future research to consider more possible factors in designing multimodal stimuli and adopt the critical criteria regarding test reliability for further advancement in the experimental approach.

CRediT authorship contribution statement

Haoyu Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Sabine van der Ham:** Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Bart de Boer:** Supervision, Resources, Methodology, Funding acquisition, Data curation, Conceptualization. **Louisa Bogaerts:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Limor Raviv:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Haoyu Zhou and Louisa Bogaerts received funding within the framework of the Odysseus programme from the Research Foundation Flanders, Belgium (FWO; project number: G0F3121N).

Sabine van der Ham and Bart de Boer were funded by the ERC project ABACUS (grant number 283435). Furthermore, Bart de Boer received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

Appendix A:. Models

Eq. (A.1). Categorization accuracy.

Categorization Accuracy ~ Condition + (1 + Condition|Participant) + (1 + Condition|Duration).

	Estimate	SE	df	t-value	p-value
(Intercept)	0.81	0.03	12.52	28.39	***<0.001***
Auditory-Non-Linguistic vs. Linguistic	0.01	0.01	118.00	1.00	0.32
Visual-Temporal vs. Spatial	−0.07	0.02	117.43	−3.83	***<0.001***
Visual-Temporal vs. Auditory	−0.06	0.01	117.80	−4.18	***<0.001***

Eq. (A.2). Production deviation from mode

Deviation from Mode ~ Condition**Category** + (1 + Condition**Category**|Participant).

*Pre-registration model: Deviation from mode ~ Condition + (1 + Condition|Participant).

	Estimate	SE	df	t-value	p-value
(Intercept)	142.11	6.14	118.51	23.15	***<0.001***
Auditory-Linguistic vs. Non-linguistic	31.66	10.38	117.86	3.05	***<0.01***
Visual-Temporal vs. Spatial	9.58	8.19	118.11	1.17	0.24
Visual-Temporal vs. Auditory	−233.01	9.69	115.75	−24.06	***<0.001***
Category value (‘*Short*’ vs. *‘Long*’)	−54.44	6.87	115.25	−7.92	***<0.001***
Category X Condition(*Auditory-Linguistic* vs. *Non-linguistic*)	−3.53	11.82	117.18	−0.30	0.77
Category X Condition(*Visual-Temporal* vs. *Spatial*)	−2.30	11.65	112.61	−0.20	0.84
Category X Condition(*Visual-Temporal* vs. *Auditory*)	5.51	9.71	111.78	0.57	0.57

Eq. (A.3). Gap difference

Gap Difference ~ Condition + (1|Participant).

	Estimate	SE	df	t-value	p-value
(Intercept)	54.13	6.94	112.59	7.80	***<0.001***
Auditory-Linguistic vs. Non-linguistic	4.24	11.24	347.85	0.38	0.71
Visual-Temporal vs. Spatial	2.75	11.22	348.89	0.24	0.81
Visual-Temporal vs. Auditory	−5.12	9.71	349.23	−0.53	0.60

Eq. (A.4). Relation between production deviation from mode and categorization accuracy

Deviation from Mode ~ Categorization Accuracy * Condition + (1 + Condition|Participant) +(1|Category).

	Estimate	SE	df	t-value	p-value
(Intercept)	110.49	35.25	20.54	3.13	***<0.01***
Categorization Accuracy	4.83	35.59	248.22	0.14	0.89
Auditory-Linguistic vs. Non-linguistic	−49.16	91.65	135.68	−0.54	0.59
Visual-Temporal vs. Spatial	19.39	29.04	213.74	0.67	0.51
Visual-Temporal vs. Auditory	−202.09	59.36	241.11	−3.40	***<0.001***
Categorization Accuracy X Condition (*Auditory inguistic* vs. *Non-linguistic*)	95.68	111.08	136.07	0.86	0.39
Categorization Accuracy X Condition (*Visual-Temporal* vs. *Spatial*)	−14.58	36.40	205.19	−0.40	0.69
Cateogry X Categorization Accuracy (*Visual-Temporal* vs. *Auditory*)	−35.03	72.51	246.95	−0.48	0.63

Eq. (A.5). Relation between production category difference and categorization accuracy

Category Difference ~ Categorization Accuracy * Condition +(1|Participant).

	Estimate	SE	df	t-value	p-value
(Intercept)	−153.72	32.18	471.01	−4.78	***<0.001***
Categorization Accuracy (*mean per pp*)	259.37	38.73	463.60	6.70	***<0.001***
Auditory-Linguistic vs. Non-linguistic	−155.45	94.15	380.74	−1.65	0.10
Visual-Temporal vs.Spatial	−44.66	60.77	415.42	−0.74	0.46
Visual-Temporal vs. Auditory	−183.70	64.22	416.06	−2.86	***<0.01***
Categorization Accuracy X Condition (*Auditory-Linguistic* vs. *Non-linguistic*)	193.41	114.19	380.92	1.69	0.09
Categorization Accuracy X Condition (*Visual-Temporal* vs. *spatial*)	92.50	74.96	416.79	1.23	0.22
Categorization Accuracy X Condition (*Visual-Temporal* vs. *Auditory*)	246.70	79.83	414.93	3.09	***<0.01***

Appendix B:. Correlation scatter plots

Table B1

Numbers of bivariate outliers removed for each correlation test using the Mahalanobis-distance method with a breakdown point of 0.25 (outliers_mcd function in the Routliers library, Leys et al., 2018).

Taskmeasures	Correlationtests					
	Auditory Linguistic ~ Auditory Non-Linguistic	Auditory Linguistic ~ Visual Temporal	Auditory Linguistic ~ Visual Spatial	Auditory Non-Linguistic ~ Visual Temporal	Auditory Non Linguistic ~ Visual Spatial	Visual Temporal ~ Visual Spatial
Categorization Accuracy	9	17	14	25	17	25
Deviation from Mode	9	10	11	7	8	13
Category Difference	15	18	19	19	18	24

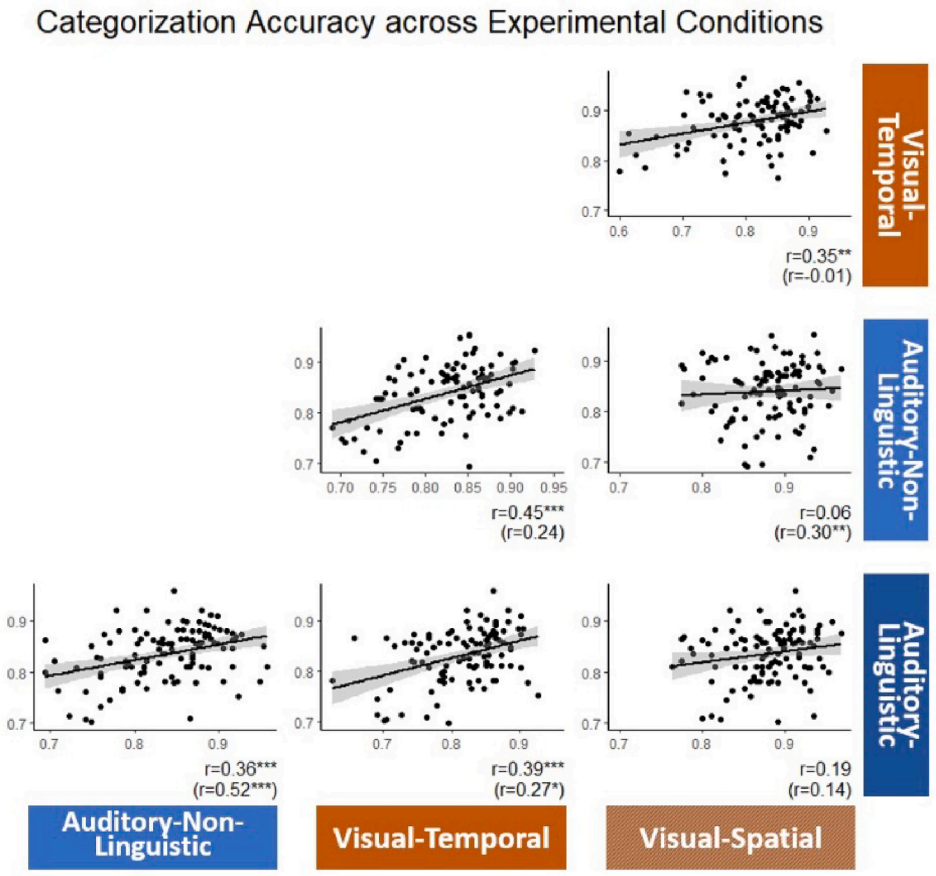

Fig. B1. Correlation matrix for categorization accuracy across experimental conditions. The scatter plots are based on the results after bivariate outlier removal. Both r-values with outlier removal and original r values (presented in parenthesis) are provided. All results were corrected using the Bonferroni test. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$.

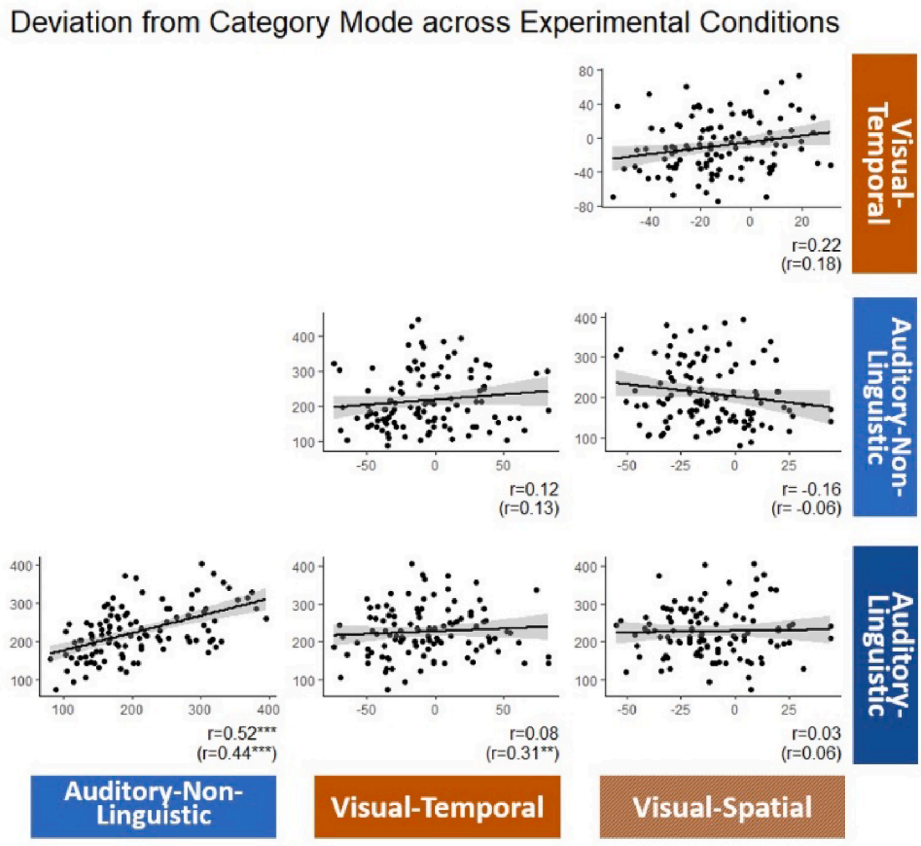

Fig. B2. Correlation matrix for deviation from mode across experimental conditions. The scatter plots are based on the results after bivariate outlier removal. Both r-values with outlier removal and original r values (presented in parenthesis) are provided. All results were corrected using the Bonferroni test. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$.

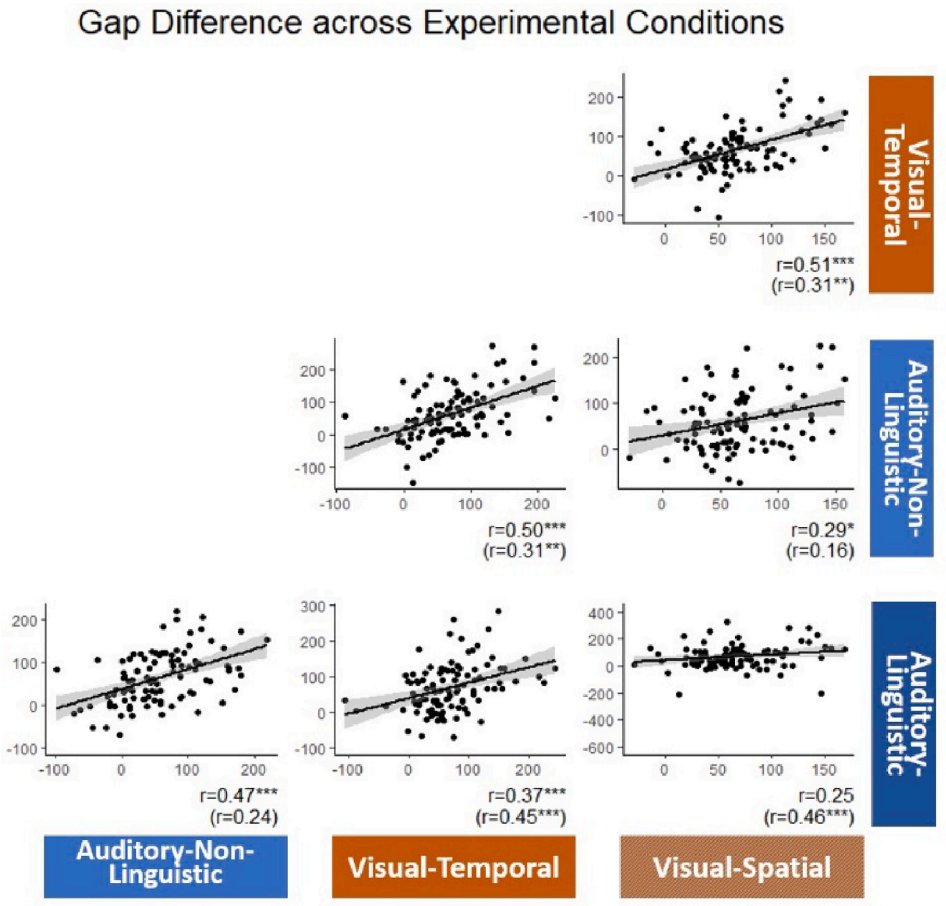

Fig. B3. Correlation matrix for gap difference across experimental conditions. The scatter plots are based on the results after bivariate outlier removal. Both r-values with outlier removal and original r values (presented in parenthesis) are provided. All results were corrected using the Bonferroni test. Correlation significance is marked with "*" as $p < 0.05$, "**" as $p < 0.01$", and "***" as $p < 0.001$.

Appendix C:. The scoring function

The problem of quantifying the participants' categorization behavior of stimuli is comparable to quantifying the quality of models that predict the probability of an outcome. An example would be meteorological models that generate the probability that it is going to rain the next day. Our participants need to specify their confidence that a stimulus falls into a certain category, but the stimuli themselves may be ambiguous. Hence for some stimuli the best thing participants can do is to provide a probabilistic statement, i.e., that the stimulus is likely to belong to a category. This is very similar to what models predicting probabilities do.

Measures to evaluate probabilistic predictors are called scoring functions, and an example that is applicable in case probabilities can be zero is the Brier score, B. For binary events where there are two possible outcomes (such as in our experiment where a category is either long or short) called E and not E, it is defined here as follows:

$$B = 1 - 1/N \sum_{i=1}^{N} (\pi_i + e_i)^2 \tag{B1}$$

where N is the number of events that is predicted, π_i is the predicted probability of event i having outcome E, and e_i is the outcome of event i, defined to be 1 if the outcome was indeed E and 0 if the outcome was not E. The Brier score is a proper score in the sense that if the predicted probabilities are the correct ones, it results in the best possible score.

Scoring functions are generally used to evaluate actual predictions, our aim is slightly different: we want to evaluate the quality of the judgment of participants about the category a stimulus represents. Mathematically, this is equivalent to how good participants' predictions would be when making repeated predictions about the category when observing a stimulus. We therefore assume N is large, and that in $p\,(l\,|s)\,N$ cases the stimulus s represents the'long' category and in $(1 - p\,(l\,|s))\,N$ cases it represents the 'Short' category. From Fig. 1 it can be seen that this probability can take on the value 0, 1/6, 1/2, 5/6 and 1. The different judgments from which the participants can choose can be approximately translated into predicted probabilities π_j (where j can take on the value of each judgment). In order to make the discussion concrete we can set these probabilities to (1, 0.8, 0.6, 0.4, 0.2, 0) in order of certainty of the judgment being that the stimulus represents the long category. The equation for the Brier score then takes the following form:

$$B_{s,j} = 1 - p(l|s)\left(1 - \pi_j\right)^2 + (1 - p(l|s)\,)\pi_j^2 \tag{B2}$$

for each combination of stimulus s and judgement j.

The actual values are then as follows:

Table C1
Illustration of actual Brier score values for stimulus categorization.

		Probability of Stimulus being long $p(l\|s)$				
		0	1/6	1/2	5/6	1
Participant Response	**Assigned probability** π_j					
Definitely Long	1	0	0.17	0.50	0.83	*1*
Probably Long	0.8	0.36	0.46	0.66	*0.86*	0.96
Maybe Long	0.6	0.64	0.67	*0.74*	0.81	0.84
Maybe Short	0.4	0.84	0.81	*0.74*	0.67	0.64
Probably Short	0.2	0.96	*0.86*	0.66	0.46	0.36
Definitely Short	0	*1*	0.83	0.50	0.17	0

where the highest scores for each stimulus are given in italics, highlighting that it is indeed a proper score (i.e., the judgment that is closest to the actual probability receives the highest score).

Although the Brier score would be usable, we felt it was unsatisfactory on two accounts: Firstly, it does not take into account that participants sometimes cannot do better than saying that a stimulus maybe represents a category, as some stimuli can indeed represent both categories. For instance, in the case of a completely ambiguous stimulus (half of the time representing the long category, half of the time representing the short category) participants can only achieve a maximum score of 0.74. Secondly, it gives relatively high scores to judgments that have to be considered very wrong, for instance they receive a score of 0.46 for classifying as'probably long' a stimulus that in reality only represents the long category one out of six times. Although especially the second issue could be addressed by tuning the π_j and doing a non-linear scaling of the Brier score, we designed a more intuitive scoring matrix, inspired by the Brier score. We therefore used the following simplified scoring table:

Table C2
Illustration of simplified scoring values for stimulus categorization.

		Probability of Stimulus being long $p(l\|s)$				
		0	1/6	1/2	5/6	1
Participant Response	**Assigned probability** π_j					
Definitely Long	1	0	0	0.20	*1*	*1*
Probably Long	0.8	0.20	0.20	0.60	*1*	0.80
Maybe Long	0.6	0.40	0.40	*1*	0.60	0.60
Maybe Short	0.4	0.60	0.60	*1*	0.40	0.40
Probably Short	0.2	0.80	*1*	0.60	0.20	0.20
Definitely Short	0	*1*	*1*	0.20	0	0

As can be seen from the elements highlighted in italics, it also has the property of a proper score. The correlation of the Brier score table and our own scoring table is 0.88. It is true that even the current scoring function is toned done from the Brier score (i.e., 0.20 accuracy instead of 0.46 for categorizing a stimulus as "probably long" when it has 1/6 probability being a long category), it still gives advantage to participants who tend to respond with middle-of-the-road options. We thus looked at frequency distribution of participant responses per probability category and found no particular tendency of responding with the middle options (Figure C.1). This was also the case across every experimental (modality) conditions.

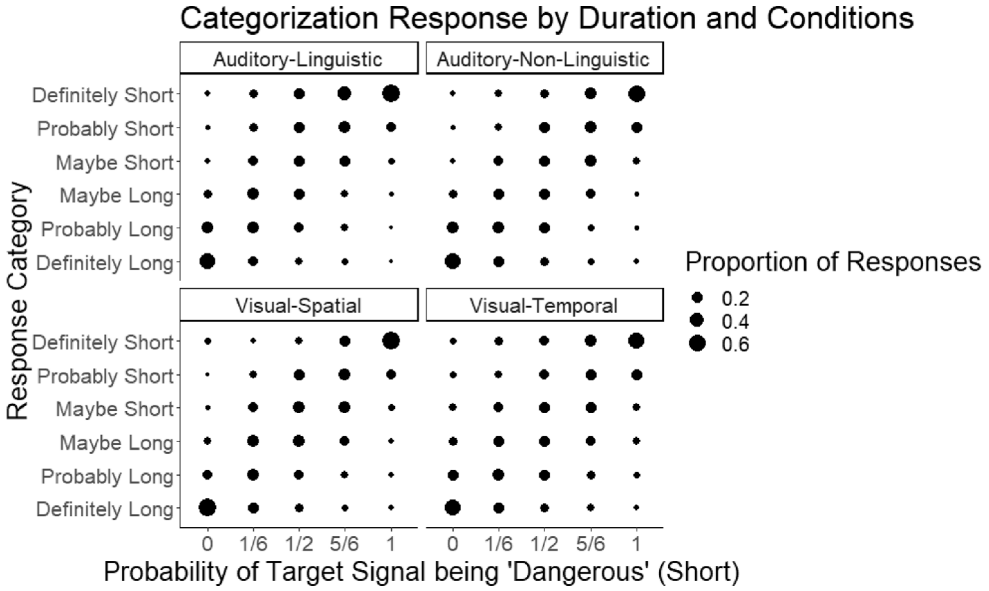

Fig. C1. The proportion of participants categorization responses by signal duration/probability of target signal being assigned to the short category across the different modality conditions.

References

Altvater-Mackensen, N., Jessen, S., & Grossmann, T. (2017). Brain responses reveal that infants' face discrimination is guided by statistical learning from distributional information. *Developmental Science, 20*(2). https://doi.org/10.1111/desc.12393

Arciuli, J., & Simpson, I. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*, 286–304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Arndt, J. (2012). Paired-associate learning. *Encyclopedia of the Sciences of Learning, 2551–2552*. https://doi.org/10.1007/978-1-4419-1428-6_1038

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models Using lme4. *ArXiv E-Prints, arXiv:1406*. doi: 10.18637/jss.v067.i01.

Boeve, S., Möttönen, R., & Smalle, E. H. (2024). Specificity of motor contributions to auditory statistical learning. *Journal of Cognition, 7*(1). https://doi.org/10.5334/joc.351

Boeve, S., Zhou, H., & Bogaerts, L. (In press). A meta-analysis of 97 studies reveals that statistical learning and language ability are only weakly correlated. *L'Année Psychologique*. doi: 10.31234/osf.io/s8mwv.

Boersma, P., & Weenink, D. (2019). *Praat: Doing Phonetics by Computer*. https://www.fon.hum.uva.nl/praat/.

Bogaerts, L., Frost, R., & Christiansen, M. H. (2020). Integrating statistical learning into cognitive science. *Journal of Memory and Language, 115*. https://doi.org/10.1016/j.jml.2020.104167

Bogaerts, L., Siegelman, N., Christiansen, M. H., & Frost, R. (2022). Is there such a thing as a 'good statistical learner'? *Trends in Cognitive Sciences, 26*(1), 25–37. https://doi.org/10.1016/j.tics.2021.10.012

Broedelet, I., Boersma, P., & Rispens, J. (2022). School-aged children learn novel categories on the basis of distributional information. *Frontiers in Psychology, 12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.799241.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87*(2), B69–B77. https://doi.org/10.1016/s0010-0277(02)00233-0

Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Attention, Perception & Psychophysics, 75*(5), 790–811. https://doi.org/10.3758/s13414-013-0475-4

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science, 11*(3), 468–481. https://doi.org/10.1111/tops.12332

Conway, C., & Christiansen, M. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology Learning, Memory, and Cognition, 31*, 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science, 17*, 905–912. https://doi.org/10.1111/j.1467-9280.2006.01801.x

Conway, C. M., & Christiansen, M. H. (2009). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology, 21*(4), 561–580. https://doi.org/10.1080/09541440802097951

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition, 170*, 312–327. https://doi.org/10.1016/j.cognition.2017.09.016

Elazar, A., Alhama, R., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the "Tabula" is anything but "rasa:" What determines performance in the auditory statistical learning task? *Cognitive Science, 46*. https://doi.org/10.1111/cogs.13102

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology (2006), 64*(5), 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning: A neural network model for statistical learning. *Cognition, 213*, Article 104621. https://doi.org/10.1016/j.cognition.2021.104621

Erickson, L., Kaschak, M., Thiessen, E., & Stutts Berry, C. (2016). Individual differences in statistical learning: conceptual and measurement issues. *Collabra, 2*, 14. https://doi.org/10.1525/collabra.41

Ferdinand, V., Thompson, B. D., Kirby, S., & Smith, K. (2013). Regularization behavior in a non-linguistic domain. *Cognitive Science*. https://www.semanticscholar.org/paper/Regularization-behavior-in-a-non-linguistic-domain-Ferdinand-Thompson/fdf3f3da8af0faae60b3c973f15f619f386c0cb6.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Fiser, J., & Lengyel, G. (2022). Statistical learning in vision. *Annual Review of Vision Science, 8*, 265–290. https://doi.org/10.1146/annurev-vision-100720-103343

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences, 105*(37), 14222–14227. https://doi.org/10.1073/pnas.0806530105

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in cognitive sciences, 19*(9), 524–533. https://doi.org/10.1016/j.tics.2015.07.009

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics, 72*(3), 561–582. https://doi.org/10.3758/APP.72.3.561

Growns, B., Siegelman, N., & Martire, K. A. (2020). The multi-faceted nature of visual statistical learning: Individual differences in learning conditional and distributional regularities across time and space. *Psychonomic Bulletin & Review, 27*(6), 1291–1299. https://doi.org/10.3758/s13423-020-01781-0

Harrell, F., & Dupont, C. (2023). *Hmisc: Harrell Miscellaneous* (5.1-0) [Computer software]. https://cran.r-project.org/web/packages/Hmisc/index.html.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research, 23*(1), 65–94.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 815–821. https://doi.org/10.1037/a0015097

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Iivonen, A., & Harnud, H. (2005). Acoustical comparison of the monophthong systems in Finnish, Mongolian and Udmurt. *Journal of the International Phonetic Association, 35*(1), 59–71.

Johnson, T., Siegelman, N., & Arnon, I. (2020). Individual Differences in learning abilities impact structure addition: better learners create more structured languages. *Cognitive Science, 44*(8), e12877.

Jung, Y., Walther, D. B., & Finn, A. S. (2021). Children automatically abstract categorical regularities during statistical learning. *Developmental Science, 24*(5). https://doi.org/10.1111/desc.13072

Kidd, E. (2012). Individual differences in syntactic priming in language acquisition. *Applied Psycholinguistics, 33*(2), 393–418. https://doi.org/10.1017/S0142716411000415

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35–B42. https://doi.org/10.1016/s0010-0277(02)00004-5

Krogh, L., Vlach, H. A., & Johnson, S. P. (2012). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology, 3*, 598. https://doi.org/10.3389/fpsyg.2012.00598

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93–107. https://doi.org/10.3758/BF03212211

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

Liu, L., Lai, R., Singh, L., Kalashnikova, M., Wong, P. C. M., Kasisopa, B., … Burnham, D. (2022). The tone atlas of perceptual discriminability and perceptual distance: Four tone languages and five language groups. *Brain and Language, 229*, Article 105106. https://doi.org/10.1016/j.bandl.2022.105106

Lukics, K. S., & Lukács, Á. (2022). Modality, presentation, domain and training effects in statistical learning. *Scientific Reports, 12*(1), Article 1. https://doi.org/10.1038/s41598-022-24951-7

Mani, N., & Schneider, S. (2013). Speaker identity supports phonetic category learning. *Journal of Experimental Psychology. Human Perception and Performance, 39*(3), 623–629. https://doi.org/10.1037/a0030402

Maye, J., & Gerken, L. (2011). Learning Phonemes Without Minimal Pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*, *24*.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3

Milne, A. E., Petkov, C. I., & Wilson, B. (2018). Auditory and Visual Sequence Learning in Humans and Monkeys using an Artificial Grammar Learning Paradigm. *Neuroscience, 389*, 104–117. https://doi.org/10.1016/j.neuroscience.2017.06.059

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62*(1), 302–331. https://doi.org/10.1111/j.1467-9922.2010.00626.x

Mitchel, A. D., Gerfen, C., & Weiss, D. J. (2016). Audiovisual perceptual learning with multiple speakers. *Journal of Phonetics, 56*, 66–74. https://doi.org/10.1016/j.wocn.2016.02.003

Ong, J. H., Burnham, D., & Stevens, C. J. (2017). Learning novel musical pitch via distributional learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(1), 150–157. https://doi.org/10.1037/xlm0000286

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, 105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

Parsons, S. (2021). splithalf: Robust estimates of split half reliability. *Journal of Open Source Software, 6*(60), 3041. https://doi.org/10.21105/joss.03041

Pavlidou, E. V., & Bogaerts, L. (2019). Implicit statistical learning across modalities and its relationship with reading in childhood. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01834

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Perfors, A., & Kidd, E. (2022). The role of stimulus-specific perceptual fluency in statistical learning. *Cognitive Science, 46*(2), e13100.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Pons, F. (2006). The effects of distributional learning on rats' sensitivity to phonetic information. *Journal of Experimental Psychology. Animal Behavior Processes, 32*(1), 97–101. https://doi.org/10.1037/0097-7403.32.1.97

R Core Team. (2021). *R: A language and environment for statistical computing [Computer software]*. R Foundation for Statistical Computing.

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21*(4), e12593. https://doi.org/10.1111/desc.12593

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81*(1), 115–119. https://doi.org/10.1037/h0027454

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews. Cognitive Science, 1*(6), 906–914. https://doi.org/10.1002/wcs.78

Rosenthal, O., Fusi, S., & Hochstein, S. (2001). Forming classes by stimulus frequency: Behavior and theory. *Proceedings of the National Academy of Sciences of the United States of America, 98*(7), 4265–4270. https://doi.org/10.1073/pnas.071525998

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110–114. https://doi.org/10.1111/1467-8721.01243

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.), 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52. https://doi.org/10.1016/s0010-0277(98)00075-4

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences, 22*(1), 52–63. https://doi.org/10.1016/j.tics.2017.10.003

Shafto, C. L., Conway, C. M., Field, S. L., & Houston, D. M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy, 17*(3), 247–271. https://doi.org/10.1111/j.1532-7078.2011.00085.x

Schapiro, A., & Turk-Browne, N. (2015). Statistical learning. *Brain mapping, 3*, 501-506. https://doi.org/10.1016/B978-0-12-397025-1.00276-1.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science, 42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), 20160059. https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Silva, S., Folia, V., Inácio, F., Castro, S. L., & Petersson, K. M. (2018). Modality effects in implicit artificial grammar learning: An EEG study. *Brain Research, 1687*, 50–59. https://doi.org/10.1016/j.brainres.2018.02.020

Smith, J. D., Berg, M. E., Cook, R. G., Murphy, M. S., Crossley, M. J., Boomer, J., … Grace, R. C. (2012). Implicit and explicit categorization: A tale of four species. *Neuroscience & Biobehavioral Reviews, 36*(10), 2355–2369. https://doi.org/10.1016/j.neubiorev.2012.09.003

Sonnweber, R., Ravignani, A., & Fitch, W. T. (2015). Non-adjacent visual dependency learning in chimpanzees. *Animal Cognition, 18*, 733–745. https://doi.org/10.1007/s10071-015-0840-x

Teinonen, T., Aslin, R., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*, 850–855. https://doi.org/10.1016/j.cognition.2008.05.009

Theeuwes, J., Bogaerts, L., & van Moorselaar, D. (2022). What to expect where and when: How statistical learning drives visual selection. *Trends in Cognitive Sciences, 26*(10), 860–872. https://doi.org/10.1016/j.tics.2022.06.001

Thiessen, E. D. (2011). Domain general constraints on statistical learning. *Child Development, 82*(2), 462–470. https://doi.org/10.1111/j.1467-8624.2010.01522.x

Thiessen, E. D., & Erickson, L. C. (2013). Beyond word segmentation: A two- process account of statistical learning. *Current Directions in Psychological Science, 22*(3), 239–243. https://doi.org/10.1177/0963721413476035

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin, 139*(4), 792. https://psycnet.apa.org/doi/10.1037/a0030801.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552. https://doi.org/10.1037/0096-3445.134.4.552

van der Ham, S., & de Boer, B. (2015). Cognitive Bias for learning speech sounds from a continuous signal space seems nonlinguistic. *I-Perception, 6*(5). https://doi.org/10.1177/2041669515593019

Vandermosten, M., Wouters, J., Ghesquière, P., & Golestani, N. (2019). Statistical learning of speech sounds in dyslexic and typical reading children. *Scientific Studies of Reading, 23*(1), 116–127. https://doi.org/10.1080/10888438.2018.1473404

Zeileis, A., & Hothorn, T. (2002). *Diagnostic Checking in Regression Relationships*. https://CRAN.R-project.org/doc/Rnews/.

Zimmerer, V., Cowell, P., & Varley, R. (2010). Individual behavior in learning of an artificial grammar. *Memory & Cognition, 39*, 491–501. https://doi.org/10.3758/s13421-010-0039-y