# Acoustic Cues Facilitate the Acquisition of Non-adjacent Dependencies in Sequences of Dynamic Object Transformation

**Zoey Zixi Lyu (ZIXILYU@Usc.Edu)**
Department of Psychology, SGM 501, 3620 S.McClintock Avenue
Los Angeles, CA 90089-1061 USA

**Neshat Darvishi (NESHATDA@Usc.Edu)**
Department of Psychology, SGM 501, 3620 S.McClintock Avenue
Los Angeles, CA 90089-1061 USA

**Toben H. Mintz (TMINTZ@Usc.Edu)**
Department of Psychology, SGM 501, 3620 S.McClintock Avenue
Los Angeles, CA 90089-1061 USA
Department of Linguistics, GFS 301, 3601 Watt Way
Los Angeles, CA 90089-1693 USA

## Abstract

Human learners' ability to detect rule-governed elements plays an important role in cognitive functions. While extensive research on the acquisition of regularities among adjacent items has provided robust and reliable evidence, learning structured patterns among non-adjacent components—known as non-adjacent dependencies (NADs)—remains far more tentative and only occurs under specific conditions. A past study by Lu and Mintz (2023) found that human learners need more training exposure to detect NADs in visual sequences of object transformations compared to sequences of human actions, but is unclear why. Building on this work, we present a series of three experiments to investigate whether learning NADs from visual dynamic sequences can be enhanced by maintaining the identifiability of the object throughout its transformations. In addition, we explore the effect of providing auditory information—speech or pure tones—along with the visual object transformation sequences. Our findings demonstrate that (a) NAD learning succeeded when speech cues co-occurred and matched with NAD-type frames but failed in the absence of auditory cues (Experiment 1); (b) pure tones presented contingently with the visual sequences also facilitated NAD learning (Experiment 2); and (c) regardless of whether speech or tones were used as additional cues, adult learners were unable to detect NADs when the relationship between specific auditory stimuli and specific visual object transformation sequences was disrupted (Experiment 3).

**Keywords:** Non-adjacent dependencies; multi-modal processing; visual statistical learning

## Introduction

Humans live in an environment saturated with events characterized by distinct regularities that require detection, extraction, and processing for learning. The ability to detect rule-governed patterns plays a crucial role in multiple domains of higher-level cognitive functions. For example, in language acquisition, recognizing the unique pattern of the present progressive tense in English grammar—where the auxiliary verb "be" precedes the present participle (e.g., verbs ending in '-ing')—is essential for constructing grammatically correct sentences to describe ongoing events. The acquisition of such regularities does not always involve conscious attention and can occur effortlessly, without cognitive control or awareness (Cleeremans, Destrebecqz, & Boyer, 1998; Dienes & Berry, 1997; Seger, 1994). A crucial cognitive skill underlying the discovery of patterns and regularities in environmental stimuli is statistical learning.

Considerable evidence demonstrates the ability to acquire regularities among adjacent items across various stimuli and sensory modalities. Human adults, infants, and even other species are able to track the distributional statistical features embedded between adjacent components in the stream of speech (Hauser, Newport, & Aslin, 2001; Romberg & Saffran, 2013; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Santolin, Rosa-Salva, Vallortigara, & Regolin, 2016; Toro & Trobalón, 2005), music tones (Saffran, Johnson, Aslin, & Newport, 1999), shapes (Fiser & Aslin, 2001, 2002), human actions (Baldwin, Andersson, Saffran, & Meyer, 2008), and emotional expressions (Mermier, Quadrelli, Turati, & Bulf, 2022). However, learning structured patterns among non-adjacent components appears to be far more challenging. Tracking pattern-based relationships between non-adjacent items is computationally more complex than tracking adjacent dependencies; it makes greater demands on working memory and requires attending to the right non-adjacent positions, potentially tracking multiple different ones. Past studies have shown that learning non-adjacent dependencies (NADs) is facilitated by certain cues that highlight the elements involved in the NAD, such as pauses between triplets (Peña, Bonatti, Nespor, & Mehler, 2002), rhythmic features that serve as bracketing cues (Wang, Zevin, & Mintz, 2017, 2019), and perceptual similarities between dependent units (Creel, Newport, & Aslin, 2004; Gebhart, Newport, & Aslin, 2009; Morgan, Meier, & Newport, 1987, 1989; Onnis, Monaghan, Richmond, & Chater, 2005; Weyers & Mueller, 2022).

This study examines visual NAD learning. In particular, we build on prior work showing cases of success and failures in visual NAD learning (Lu & Mintz, 2023), and explore two kinds of information that could enhance NAD learning: 1) visual information that could help "packaging" the sequences in a way that makes NADs more salient, and 2) multi-modal



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

stimuli that could provide cues to categories of NADs.

### Detecting Non-Adjacent Dependencies in Visual Action Sequences

In the visual domain, structural information is present in dynamic scenes that involve sequences of continuous events. For example, human actions, like language and music, consist of unique components arranged in ordered sequences (Fitch & Martins, 2014). Previous research has provided evidence of successful NAD learning from sequences of human actions (Endress & Wood, 2011; Li & Mintz, 2015; Lu & Mintz, 2023). Furthermore, NADs appeared easier to learn from sequences of human action compared to similarly structured sequences of object transformations (Lu & Mintz, 2023). Lu and Mintz presented learners with triplets of either sequences involving human avatars, where the first and third elements of the sequence comprised an NAD, or object sequences following the same patterns. The results revealed that while NADs could be learned from both human actions and object transformations, longer training periods were required for learning from objects compared to humans. Likewise, subjects needed longer training exposure to sequences of static human postures compared to sequences of dynamic human actions. These findings suggest that stimuli involving human forms and, independently, stimuli involving dynamic motion each enhance NAD learning.

Lu and Mintz (2023) proposed two possible explanations for why learners struggled to acquire NADs from object transformations in the same amount of training time as from human actions. First, learners may benefit from the fact that human avatars are perceived as conspecific. The familiar forms could lead to easier processing and enhanced memory for the sequences. In addition, viewing human action has been shown to activate human viewers' motor systems (Reid, Kaduk, & Lunn, 2019; Salo, Ferrari, & Fox, 2019; Wilson & Knoblich, 2005), which could result in richer encoding and subsequent processing of the NADs. Second, the difficulty in learning NADs from object transformations could have resulted from the lack of identifiable features in the animated object, making it challenging for participants to perceive them as consistent entities throughout their transformations (the object in Lu and Mintz was a flat plane that dynamically transformed into different shapes). While it is not clear why maintaining the identity of the object across the three transformations would be critical, it is possible that in so doing, learners process the sequence of three transformations more as cohesive unit, thus bringing the first and last element—the critical elements for the NADs—under the same representation. In order to test this possibility, in this study we used an object with clearly identifiable features that were maintained across all transformations.

### The Influence of Language in Object Categorization

In the field of child word learning, extensive research has demonstrated the relationship between naming and object categorization (Fulkerson & Haaf, 2003; Waxman & Markow, 1995; Waxman & Braun, 2005) and individualization (Xu, 2002; Xu, Cote, & Baker, 2005). Moreover, object naming facilitates the establishment of object categorization and conceptual hierarchies. For example, infants group objects into distinct categories when those objects are referred to by different words (Dewar & Xu, 2007; Plunkett, Hu, & Cohen, 2008; Waxman & Braun, 2005). An infant study performed by LaTourrette and Waxman (2020) further supports the role of object naming in object representation. In that study, infants were assigned to one of the three conditions in which four novel objects were presented in a row, accompanied by either distinctive nouns (distinctive name condition), identical nouns (consistent name condition), or tone sequences (control condition). Later, infants could reliably recognize a previously displayed object when shown alongside a new object from the same category in the distinctive name condition. These findings strongly support the claim that labeling contributes to object encoding and recalling. Furthermore, they cast light on the underlying mechanisms by which learners build concepts on objects, pointing in particular to the different guiding effects of labels: consistent labels highlight commonalities between objects and inconsistent labels highlight distinctiveness rooted in objects. Altogether, these insights suggest that pairing auditory labels with triplets of object transformations, as in Lu and Mintz (2023), may similarly facilitate learning by making the NAD rules in dynamic visual input more salient through labeling.

Building on the aforementioned studies, we investigated whether NAD patterns could be detected in object transformations when the object could be easily identifiable as a consistently well-maintained entity over time. Moreover, we examined whether auditory labels paired with object transformation sequences—akin to linguistic labels—would facilitate the acquisition of NAD rules.

## Experiment 1: Silence and Speech

In Experiment 1, we created a 3-D torus object with three spikes, as shown in Figure 1. During object transformations, the spikes moved as part of the object, which was expected to help participants track the identity of the object. The experiment included two conditions: a visual-only condition and a synthesized speech condition. In the synthesized speech condition, additional auditory labels for NAD frames were provided using synthesized human speech. If adding matched speech facilitates NAD detection and participants learn the NADs in both conditions, we expected a greater effect size in the speech condition compared to the silent condition. Alternatively, if no learning occurred in the silent condition, learning in the speech condition would provide evidence that the correlated auditory information facilitated the detection of NADs.

### Methods

**Participants** We recruited 172 students from the undergraduate population at the authors' institution. Recruitment

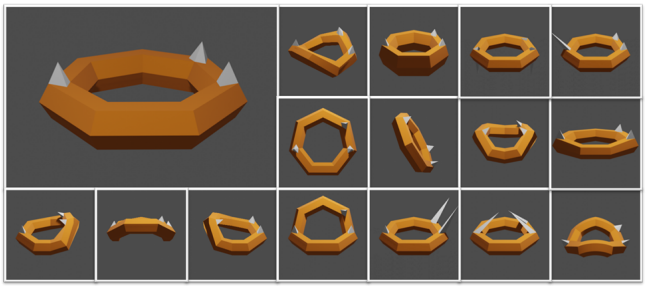

Figure 1: The large image depicts the object in its neutral position. The smaller images depict the maximum transformation frame of the 15 distinct transformations.

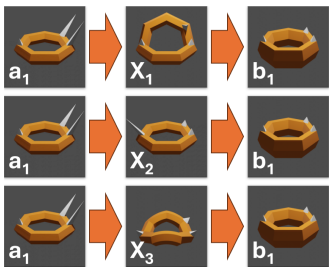


Figure 2: Example of the three triplets from the $a_1\_b_1$ NAD frame, including $a_1X_1b_1$, $a_1X_2b_1$, and $a_1X_3b_1$. The triplets play from left to right in 1.875 seconds accompanied with sentence "Look, it's huring". Each image depicts the midpoint (maximum) of the dynamic motion from the neutral form and back that constitutes the actual video clips.

was conducted through the SONA subject pool, and participants received course credit as compensation. We excluded 70 participants from the analysis: 64 due to failure in the attention checks (see Procedure for details) and 6 due to technical issues. The final sample includes 102 participants, where 55 of them are assigned to visual-only condition and the other 47 are assigned to synthesized speech condition. The exclusion criteria and rigorous screening process ensured that the final sample comprised attentive and engaged participants, enhancing the reliability and validity of the study's results.

**Materials** The visual stimuli consisted of 15 video clips depicting the dynamic transformations of a 3D torus with three distinct spikes. The large image in Figure 1 depicts the object in its neutral position. Each video illustrated a single action (e.g., bending or expanding along a horizontal axis) lasting 625 milliseconds. The smaller images in Figure 1 depict the most extreme degree of transformation from the neutral position, which occurred midway in the 625 ms clip. Following each action, the torus returned to its neutral position, providing clear boundaries between actions and facilitating sequence segmentation. These video clips were developed using Blender, a 3D computer graphics software (Community, 2018) and were used for both the training and testing phases.

Training Stimuli For each participant, the 15 video clips were randomly assigned to $a_1$, $a_2$, $a_3$, $b_1$, $b_2$, $b_3$, $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, $X_6$, $X_7$, $X_8$, $X_9$. Then they were organized into three non-adjacent dependency (NAD) frames: $a_1\_b_1$, $a_2\_b_2$, and $a_3\_b_3$. Each frame featured a co-dependent first ($a$) and third ($b$) action, with a variable intervening action ($X$). This arrangement resulted in nine unique NAD triplets (i.e., $a_1X_1b_1$, $a_1X_2b_1$, $a_1X_3b_1$, $a_2X_4b_2$, $a_2X_5b_2$, $a_2X_6b_2$, $a_3X_7b_3$, $a_3X_8b_3$, $a_3X_9b_3$) for the training phase. Figure 2 shows examples of the three $a_1\_b_1$ NAD frame triplets, where the three triplets share the first action and the last action.

The auditory stimuli used in this experiment were three synthesized speech phrases ("Look, it's huring," "Look, it's daping," and "Look, it's toving"). The speech was synthesized using the pyttsx3 package in Python 3.11 with Microsoft Zira voices. The speech was generated with a speech rate of 150 and a medium pitch level of 10. The speech was adjusted to the same length as the video triplet by padding at the end and was played simultaneously with the video. The three sentences were paired, one-to-one, to the three NAD frames throughout the training phase.

Testing Stimuli The test triplets included 18 novel NAD-test-triplets that adhered to the NAD structures established during training but introduced new intervening elements (i.e., $a_1X_4b_1$, $a_1X_5b_1$, ..., $a_1X_9b_1$, $a_2X_1b_2$, $a_2X_2b_2$, $a_2X_3b_2$, $a_2X_7b_2$, $a_2X_8b_2$, $a_2X_9b_2$, $a_3X_1b_3$, $a_3X_2b_3$, ..., $a_3X_6b_3$). These test sequences were novel in the sense that the transitional probability between each action clip was 0, however, the sequences maintained the NAD relationships between the first and third action clips that occurred in the training stimuli. There were also 18 positional-test-triplets that were created by taking the first and third action from two different training NAD frames and introducing an intervening action that did not occur with either edge element during training (i.e., $a_1X_{4-6}b_3$, $a_1X_{7-9}b_2$, $a_2X_{1-3}b_3$, $a_2X_{7-9}b_1$, $a_3X_{1-3}b_2$, $a_3X_{4-6}b_1$). As with the NAD-test-triplets, the transitional probability between each action clip was also 0, but unlike the NAD-test-triplets, the positional-test-triplets did not maintain the dependency relationships that occurred in the familiarization stimuli. The positional-test-triplets thus match all properties of the NAD-test-triplets, except they violate the trained NADs.

An additional 12 "catch" trials featured entirely novel and repeated actions using only $a$ and $b$ that were not in the same NAD frame (e.g., $a_1b_2b_2$ or $a_1a_1b_2$). They should have been easily identified as novel sequences, as training sequences never contained repetitions and items occur in sequential position where they didn't occur in the training material. All the test stimuli were presented without auditory stimuli, regardless of whether they were for the visual-only condition or synthesized speech condition.

**Testing Environment** Participants completed the experiment using their own laptops or desktops. Mobile devices were not allowed for the experiment. Before the experiment, they confirmed that they had normal or corrected-to-normal vision and hearing. A simple hearing test that played a syn-

thesized voice "cat" and asked the participant to input what they had heard was conducted to ensure they could perceive the auditory stimuli. Participants were also required to confirm that they were viewing the experiment in full-screen mode before proceeding.

After the experiment, participants were asked whether they encountered any technical issues. Data from those who reported technical issues were excluded from the analysis.

**Procedure** During the training phase participants were exposed to 180 triplets presented in a pseudo-randomized order. These triplets were constructed from the nine unique NAD triplets described earlier, and each triplet repeated 20 times. Each triplet lasted 1875 milliseconds (3 actions times 625 milliseconds per action) plus a 125-millisecond pause between triplets. Participants in the visual-only condition viewed silent triplets, while those in the speech condition additionally received auditory cues. The training phase lasted approximately 6 minutes.

The testing phase evaluated participants' ability to recognize NAD triplets and distinguish them from positional triplets. Recall that both types of test items were novel sequences in which each component action was in the same position in which it occurred during training, but only the NAD triplets adhered to the NAD structure of the training stimuli. Participants were presented with 48 test triplets and were asked to rate the familiarity after each triplet on a 5-point scale, ranging from 1 ("definitely had not seen") to 5 ("definitely had seen"). The triplets were presented in a pseudo-randomized order to prevent patterns in presentation from influencing responses, with no more than two triplets of the same type appearing consecutively. However, the initial 30 participants in the synthesized speech condition were shown randomly ordered sequences accidentally.

Although both test sequence types were novel, each action occurred in its trained position within the triplet, so we expected learners to rate all items as somewhat familiar. However, if they detected the NADs, they should rate the NAD items higher, since the identical NAD patterns occur in these items compared to the training triplets.

## Results and Discussion

Before the data were analyzed, participants who rated 4 or more catch trials with a familiarity score of 3 or higher were excluded from further analysis. We then fitted an ordinal logistic regression model using the "ordinal" package in R (Christensen, 2019) to investigate the relationship between participants rating and the type of the stimuli. The model includes the fixed effect of the test trial number, study condition, and the interaction between the study condition and the stimuli type. In addition, participant-level random effects for the stimuli type and trial number are also included. In the visual-only condition, the participants' average rating was 3.28±0.52 for NAD triplets and 3.26±0.45 for positional triplets, and in the speech condition, the participants' average rating was 3.23±0.60 for NAD triplets and 3.09±0.61 for positional triplets. Figure 3 shows the box plot of the average difference between the scores in NAD triplets and positional triplets. The model indicates that in the visual-only condition, there was no evidence that the ratings for the NAD and positional triplets were different ($\beta$=0.035, z=0.821, p=0.411). However, in the speech condition the difference between the ratings for the NAD and positional triplets significantly differed from 0 ($\beta$=0.118, z=2.466, p=0.014). There was also a significant effect from the test trial number ($\beta$=-0.014, z=-3.754, $p<0.01$), showing that the participants tend to give lower scores as the test phase went on.

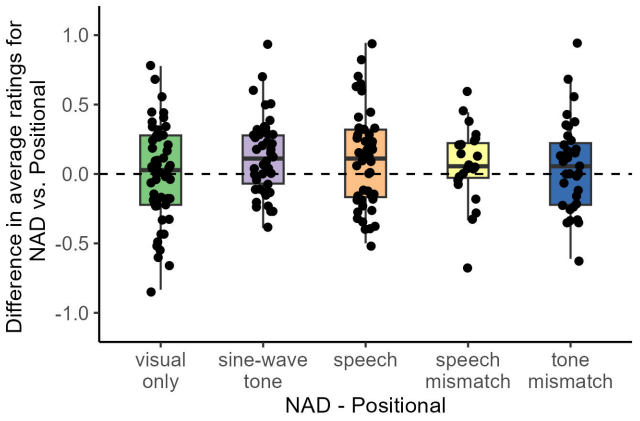


Figure 3: Box plot of the average difference in scores for NAD triplets versus positional triplets, Experiments 1–3. Each black dot represents a participant.

Consistent with the findings of Lu and Mintz (2023), these results suggest that human adults were unable to learn NADs of object transformations after six minutes of exposure, even when the identity of the object is clearly maintained during the transformations. However, the results show that they were able to learn NADs when they also heard speech stimuli during the exposure phase, and when each label used was consistently paired with a specific NAD. The findings suggest that maintaining the identity of the object over transformations may not be the primary barrier to NAD learning from objects. Instead, successful learning from human avatars may be attributed to the distinctive characteristics of the avatars, i.e., human participants can perceive the actions more efficiently as they can perform similar actions. The results in the speech condition indicate that correlated auditory cues facilitates NAD learning, even when learning from objects is inherently challenging. However, since this experiment utilized human speech, it remains unclear whether the communicative nature of speech sounds is the key factor driving this success.

# Experiment 2

Experiment 1 demonstrated that adults can learn NADs from sequences of dynamic object transformations when each type of NAD was paired with unique auditory information from co-occurring speech. Experiment 2 aimed to further investi-

gate whether the linguistic properties of the auditory information played a critical role. Specifically, much like linguistic labels have been shown to facilitate object categorization and generalization in children (LaTourrette & Waxman, 2020; Waxman & Markow, 1995; Waxman & Braun, 2005), it is possible that the novel word in the co-occurring linguistic phrase—e.g., "Look, it's *toving*"—was treated as a label for the triplet action sequence it occurred with. In that case, even though there were three distinct sequences that each particular novel word occurred with—each differing in the middle item—their occurrence with the same word could have caused them to notice the similarity between the sequences, namely, the NAD. On the other hand, the fact that the auditory stimuli were speech and that they contained potential labels may have been irrelevant to their facilitatory effect. Rather, it could have been that any auditory information that was consistent across sequences that contained the same NAD was responsible for the enhanced NAD learning in the speech condition in Experiment 1. To test this, in Experiment 2, we replaced the speech stimuli with a pure sine-wave tone. Similar to Experiment 1, a distinct tone was paired with a specific NAD. Previous studies by Ferry, Hespos, and Waxman (2010) have shown that tones alone, unless perceived as communicative, do not facilitate word learning in infants. However, there is limited evidence regarding whether this finding extends to adults. In this experiment, we removed the communicative property of the auditory information by replacing speech with non-communicative sine-wave tones.

### Methods

**Participants** Another 90 undergraduate students from the authors' institution who had not participated in Experiment 1 were recruited through the SONA subject pool with the same compensation. A total of 44 participants were excluded from the analysis: 40 due to failed attention checks and 4 due to technical issues, resulting in a final sample of 46 participants.

**Materials** The materials in Experiment 2 were the same as Experiment 1's synthesized speech condition except that the synthesized speech phrases were replaced by sine-wave tones of 300Hz, 350Hz, and 400Hz. The sine-wave tones were generated by scipy package in Python 3.11. Here, instead of pairing a unique word with a specific NAD, a specific tone was paired consistently with a specific NAD.

**Procedure** The procedure for Experiment 2 was identical to the speech condition in Experiment 1, except that the synthesized speech phrases were replaced by sine-wave tones. As before, no auditory information accompanied the test triplets.

### Results and Discussion

The same model as in Experiment 1 was fitted for Experiment 2. The participants' average rating was 3.35±0.46 for NAD triplets and 3.17±0.50 for positional triplets, and the rating for the NAD was significantly larger than that of positional triplets ($\beta$=0.130, z=2.874, $p<0.01$), suggesting that participants learned the NAD patterns in the training materials. A significant main effect was also found for test trial number ($\beta$=-0.027, z=-4.711, $p<0.01$). These results demonstrate that human adults can learn NADs in sequences of dynamic object transformations not only with accompanying speech that varies consistently with NAD type, but also when the visual NAD sequences were accompanied by consistently varying pure tones. This suggests that the communicative nature of the auditory information is not necessary for facilitating NAD learning. While it can be argued that the auditory labels provided additional representativeness to the object transformation sequences, we have no evidence that this is the case. Rather, it appears that providing different types of auditory information, by providing multi-modal correlated cues to NAD identity, facilitated learning the NAD sequences.

In sum, Experiments 1 and 2 demonstrate that when a novel object was viewed undergoing a sequence of dynamic transformations involving a non-adjacent dependency structure, the visual information alone was not sufficient to facilitate NAD learning with 6 minutes of exposure, as shown in prior experiments by Lu and Mintz (2023). However, under the same visual experience, adding a broad range of correlated auditory stimuli during the learning phase facilitated NAD encoding.

## Experiment 3

In Experiments 1 & 2, the auditory information was always consistently paired with NAD type. If learners were using the auditory information as a cue to aid in detecting the similarities in the visual sequences—i.e., the NADs—then providing the same auditory information, but uncorrelated with NAD type, should not lead to such facilitation. Alternatively, the auditory stimuli could simply have played a more general role, such as by increasing overall attention to the experimental task. In that case, learners might benefit from multi-modal stimuli even if it does not correlate with the NAD type. To investigate this, in Experiment 3 we retained the auditory stimuli used in Experiments 1 and 2 but disrupted the matching between the auditory sequences and the NAD frames. This manipulation ensured that the auditory cues could no longer serve as correlated cues for the NAD frames.

### Methods

**Participants** Another 98 undergraduate students from the authors' institution who had not participated in either Experiment 1 or Experiment 2 were recruited through the SONA subject pool with the same compensation. A total of 41 participants were excluded from the analysis: 36 due to failed attention checks and 5 due to technical issues, resulting in a final sample of 57 participants, where 24 participants were assigned to the speech condition and 33 participants were assigned to the sine-wave tone condition.

**Materials** The visual training and test stimuli and design were the same as in Experiments 1 and 2. The auditory materials were the same as in the relevant conditions in Experiments 1 and 2, except that the auditory information now

mismatched the NAD frames. For example, "Look, it's hurring" now occurred with three visual sequences that contained different NADs, for example $a_1X_1b_1$, $a_2X_4b_2$, and $a_3X_7b_3$. In this way, each word/tone was paired with all three NAD frames, and no word/tone was paired in any privileged way with a particular NAD.

**Procedure** The procedure for Experiment 3 was identical to Experiment 2 and the speech condition in Experiment 1, except that the auditory labels were misaligned with the NAD frames.

### Results and Discussion

The same model used in Experiments 1 and 2 was fitted. In the speech condition, the participants' average rating was 3.46±0.37 for NAD triplets and 3.34±0.44 for positional triplets, and in the sine-wave tone condition, the participants' average rating was 3.27±0.54 for NAD triplets and 3.20±0.54 for positional triplets. The model indicated that in both conditions, the ratings for the NAD and positional triplets were not significantly different (speech: $\beta$=0.066, z=1.083, p=0.279; sine-wave tone: $\beta$=0.066, z=1.205, p=0.228). A significant effect was found for trial number ($\beta$=-0.019, z=-3.610, $p<0.01$). The results of the model showed no evidence of learning the visual NADs when the auditory information mismatched the visual NAD frames, regardless of the type of auditory stimuli. This suggests that multimodality alone does not facilitate NAD learning; instead, the consistent pairing of the auditory information with the visual NAD frames appears to be a prerequisite for successfully encoding the NADs under the general exposure conditions across the three experiments.

## General Discussion

In this study, we investigated whether adult learners could identify NAD patterns from the transformations performed by an identifiable object after six minutes of training exposure. We also examined the role of auditory labels in NAD learning, specifically focusing on the effects of communicative and non-communicative cues when paired congruently or incongruently with NAD-type frames. Our findings revealed that adults struggled to learn NADs from object transformations without auditory cues, even when the object remained clearly identifiable as the same object throughout the transformation sequence. Similarly, learning failed when auditory cues did not match the NAD triplets. However, under comparable exposure conditions, we found that adults could successfully learn NADs from dynamic object transformations when the auditory cues corresponded to the NAD-type frames.

Our study replicates and extends the findings of Lu and Mintz (2023), who demonstrated that dynamic motions yield stronger facilitating effects in human agents than in nonhuman agents. Specifically, Lu and Mintz (2023) found that adults could learn NADs from action sequences performed by human avatars but struggled to learn from object transformations without extended training. They attributed the failure to the shape of the object (a flat plane) that the object's persistent variation in configuration undermined its consistent shape attributes, making it harder for the participants to identify the object as a coherent entity rather than multiple distinct objects during transformations. In our study, we addressed this limitation by designing a three-dimensional object with prominent and stable features to help participants perceive and encode it as a unified whole. Nevertheless, under the same duration of exposure (i.e., six minutes) as in Lu and Mintz (2023), we did not find evidence of NAD learning. Although our object with its distinctive features is arguably easier to track across transformations than a flat plane (Lu & Mintz, 2023), the transformations themselves may be less distinctive from one another, leading to an increased difficulty of identifying the NAD patterns. The detection of NADs from the object transformation involves two key processes: tracking a variety of unique actions that the object undergoes, and extracting and summarizing the abstract rules across training triplets. The failure to learn NADs in the silent condition may have been due to difficulty in tracking the actions. However, in the conditions when the tones and speech were correlated with the NADs, the auditory information could have provided a grouping cue that helped learners notice the similarities in the triplets with the same NADs, and thereby make NAD-based generalizations that result in ranking the NAD-conforming test items higher than the ones with mere position similarity.

Our findings that adult learners could acquire visual NADs with corresponding auditory cues underscore the significance of correlated cues in extracting statistical regularities (Gerken, Wilson, & Lewis, 2005). More importantly, our findings indicate that tones, similar to speech, support NAD learning. Previous studies on abstract rule learning from tones attribute the successful learning to the communicative purpose that the tone carries, when the tones are given communicative functions (Ferguson & Lew-Williams, 2016). However, lacking a communicative function, the tones in our study did not serve as communicative signals. Instead, we speculate that, as with correlated speech information, they served as a categorization cue that helped participants to notice the visual regularities that constituted the NAD patterns. The distinct auditory cues that correspond to an NAD frame highlight the visual similarities among triples from the same NAD frame and the visual differences among triplets from different NAD frame, aiding the comparison process. Our study adds to the strands of previous investigations that confirm the boosting impact of constraints on NAD learning. In particular, our finding shows that the corresponding relationship between auditory information and the visual stimuli could help specify the range of the sequences where NAD is computed, resulting in an improved learning as same as other perceptual cues that emphasize the structured relationship (Endress, Nespor, & Mehler, 2009; Grama, Kerkhoff, & Wijnen, 2016; Newport & Aslin, 2004; Wang et al., 2019).

## References


Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407.

Christensen, R. H. B. (2019). Ordinal—regression models for ordinal data. *R package version*, *10*(2019), 54.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in cognitive sciences*, *2*(10), 406–416.

Community, B. O. (2018). Blender - a 3d modelling and rendering package [Computer software manual]. Stichting Blender Foundation, Amsterdam.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, memory, and cognition*, *30*(5), 1119.

Dewar, K., & Xu, F. (2007). Do 9-month-old infants expect distinct words to refer to kinds? *Developmental psychology*, *43*(5), 1227.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic bulletin & review*, *4*, 3–23.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in cognitive sciences*, *13*(8), 348–353.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive psychology*, *63*(3), 141–171.

Ferguson, B., & Lew-Williams, C. (2016). Communicative signals support abstract rule learning by 7-month-old infants. *Scientific reports*, *6*(1), 25434.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3-and 4-month-old infants: an advantage of words over tones. *Child development*, *81*(2), 472–479.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*(1), 87–104.

Fulkerson, A. L., & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9-and 15-month-olds' object categorization. *Infancy*, *4*(3), 349–369.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic bulletin & review*, *16*(3), 486–490.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(2), 249–268.

Grama, I. C., Kerkhoff, A., & Wijnen, F. (2016). Gleaning structure from sound: The role of prosodic contrast in learning non-adjacent dependencies. *Journal of psycholinguistic research*, *45*, 1427–1449.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64.

LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences*, *117*(35), 21230–21234.

Li, J., & Mintz, T. H. (2015). Constraints on learning non-adjacent dependencies (nads) of visual stimuli. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 1350–1355). Austin, TX: Cognitive Science Society.

Lu, H. S., & Mintz, T. H. (2023). Dynamic motion and human agents facilitate visual nonadjacent dependency learning. *Cognitive Science*, *47*(9), e13344.

Mermier, J., Quadrelli, E., Turati, C., & Bulf, H. (2022). Sequential learning of emotional faces is statistical at 12 months of age. *Infancy*, *27*(3), 479–491.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*(4), 498-550.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1989). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language*, *28*(3), 360-374.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. statistical learning of non-adjacent dependencies. *Cognitive psychology*, *48*(2), 127–162.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225–237.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604–607.

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681.

Reid, V. M., Kaduk, K., & Lunn, J. (2019). Links between action perception and action production in 10-week-old infants. *Neuropsychologia*, *126*, 69–74.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive science*, *37*(7), 1290–1320.

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.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., &

Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, *8*(2), 101–105.

Salo, V. C., Ferrari, P. F., & Fox, N. A. (2019). The role of the motor system in action understanding and communication: Evidence from human infants and non-human primates. *Developmental psychobiology*, *61*(3), 390–401.

Santolin, C., Rosa-Salva, O., Vallortigara, G., & Regolin, L. (2016). Unsupervised statistical learning in newly hatched chicks. *Current Biology*, *26*(23), R1218–R1220.

Seger, C. A. (1994). Implicit learning. *Psychological bulletin*, *115*(2), 163.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & psychophysics*, *67*(5), 867–875.

Wang, F. H., Zevin, J., & Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cognitive Psychology*, *113*, 101223.

Wang, F. H., Zevin, J. D., & Mintz, T. H. (2017). Top-down structure influences learning of nonadjacent dependencies in an artificial language. *Journal of Experimental Psychology: General*, *146*(12), 1738.

Waxman, S. R., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. *Cognition*, *95*(3), B59–B68.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Weyers, I., & Mueller, J. L. (2022). A special role of syllables, but not vowels or consonants, for nonadjacent dependency learning. *Journal of Cognitive Neuroscience*, *34*(8), 1467–1487.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological bulletin*, *131*(3), 460.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223–250.

Xu, F., Cote, M., & Baker, A. (2005). Labeling guides object individuation in 12-month-old infants. *Psychological Science*, *16*(5), 372–377.