RUNNING HEADER: FINDING STRUCTURE IN MODERN DANCE

Finding Structure in Modern Dance

Claire Monroy[a] & Laura Wagner[b]

[a]School of Psychology, Keele University,

Dorothy Hodgkin Building, Keele ST5 5BG, United Kingdom

Email: c.d.monroy@keele.ac.uk

[b]Department of Psychology, The Ohio State University

835 Neil Avenue, Columbus, Ohio, 43210, United States

Email: wagner.602@osu.edu

Corresponding Author: Claire Monroy

Keywords: event segmentation, action sequences, modern dance, intentions, language

Abstract

Research has shown that both adults and children organize familiar activity into discrete units with consistent boundaries, despite the dynamic, continuous nature of everyday experiences. However, less is known about how observers segment *unfamiliar* event sequences. In the current study, we took advantage of the novelty that is inherent in modern dance. Modern dance features natural human motion but does not contain canonical goals—therefore, observers cannot recruit prior goal-related knowledge to segment it. Our main aims were to identify whether observers segment modern dance into the steps intended by the dancers, and what types of cues contribute to segmentation under these circumstances. Experiment 1 used a classic event segmentation task and found that adults were able to consistently identify only a few of the dancers' intended steps. Experiment 2 tested adults in an off-line labeling task. Results showed that steps which could more easily be labeled off-line in Experiment 2 were more likely to be segmented on-line in Experiment 1.

Keywords: event segmentation, action sequences, modern dance, intentions, language

1.0 Introduction

Consider a simple, everyday event, like a barista preparing a cappuccino in our favorite café. As we wait, we observe a sequence of individual steps: grinding beans, foaming milk, reaching for a cup, pouring the liquid. Though the movements are executed in a continuous, fluid stream, we segment the stream into meaningful units and perceive boundaries between the adjacent steps. Decades of research has established that *event segmentation* is a natural and automatic feature of processing continuous action sequences like this one, which are familiar and contain canonical goals (Zacks, 2004). However, what about sequences that are *unfamiliar*, featuring no obvious goals, easily labeled actions, or relevant objects? We note that for infants, novel experiences are an everyday feature of life, as infants are constantly learning about new actions, goals, and linguistic labels. The current study did not focus on infants, but instead placed adults in a comparable situation involving modern dance sequences. Modern dance is made of continuous intentional human action but is composed of movements that are novel for most observers; moreover, few of the actions result in any perceptible goal or outcome, nor do they have common labels. The current study takes advantage of this unique aspect of modern dance to investigate the role of kinematic and linguistic cues in event segmentation.

The available cues to assist in event segmentation range from perceptual cues that help us analyze actions from the bottom up, to conceptual cues that help us break things down in a top-down way. At a perceptual level, adults and infants can segment familiar and novel action sequences using broad spatiotemporal cues such as pauses (Sharon & Wynn, 1998; Wagner & Carey, 2003) and statistical regularities (Baldwin, Andersson, Saffran, & Meyer, 2008; Monroy, Gerson, & Hunnius, 2018; Stahl, Romberg, Roseberry, Golinkoff, & Hirsh-Pasek, 2014).

Another type of cue that we can leverage to find events within a continuous stream of action is our understanding of how human movement works. Human movement is constrained by physical anatomy, forces of gravity, and dynamics of motion like velocity and acceleration. For example, if a dancer jumps into the air, he must also (eventually) come back down again, ending the jumping event. Similarly, our anatomy constrains the possible range of movements that we make during everyday actions. Research on the perception of action and movement has demonstrated that, from early in life, human observers are sensitive to the kinematic and biological properties of motion (Savaki et al., 2022; Stapel, Hunnius & Bekkering, 2012). For instance, infants as young as four months of age prefer looking at biological point-light displays over randomly moving dots (Fox & McDaniel, 1982; Johansson, 1973). While infants' developing motor skills also likely influence their developing appreciation for how motion kinematics work (Adolph & Hoch, 2019; Gerson et al., 2015), these findings indicate that sensitivity to the unique information contained in biological motion emerges from the earliest moments in life and is a foundational aspect of the human visual system. Put simply, our visual system is carefully attuned to the properties of human movement.

Even young infants are also able to use more conceptual information about events to help them find meaningful units. In Hespos et al. (2009; see also Hespos, Grossman, & Saylor, 2010) 6–8 month old infants could keep track of schematic event sequences involving containment, support, and occlusion. The events themselves were presented in a continuous fashion via a bouncing ball that smoothly moved from one type of event to the other. Infants noticed deviations in the ordering of the events—even when they involved the same objects—suggesting that infants were segmenting the sequence into coherent units based on conceptual schema. In addition, there is ample evidence that infants use information corresponding to an actor's intention to segment continuous actions into events. For instance, Baldwin et al. (2001) demonstrated that 10-month-olds would look longer (i.e., show more surprise) at a simple event (a woman dropping a towel and bending down to retrieve it) if a pause was inserted that interrupted the action (e.g. while the woman was bending) compared to when the woman achieved her goal (e.g., grasping the towel). This result suggests that the infant perceived that the actor's intention was also the event's expected boundary point.

Not surprisingly, adults also use actors' intentions to identify event units. Levine et al., (2017) showed experts and novices continuous sequences of an ice-skating routine and found a strong alignment between the two groups in the identification of event boundaries within the sequence; the alignment was stronger for the goals (endpoints) of these events than the sources (starting points) of the events, highlighting the importance that adults place on intentional outcomes. That said, familiarity with an event is clearly an important factor in helping adults appropriately analyze those intentions. Levine and colleagues chose ice-skating precisely because many adults are unfamiliar with closely analyzing it; and indeed, they found that expertise with ice-skating was a better predictor than perceptual cues of participants' ability to consistently identify event boundaries. Similarly, Wurm & Schubotz (2012) found that adults were more accurate at recognizing actions when they were depicted in a familiar context (cracking an egg in a kitchen) rather than in an unfamiliar context (cracking a egg in a woodshop). Adults automatically bring their rich experiences with events with them when they analyze new scenes.

One particularly rich source of event knowledge which guides event segmentation is language (Papafragou & Selimis, 2010; Papafragou et al., 2008). Indeed, one of the concepts that is used across languages to organize verb semantics and its syntactic instantiation into sentences is telicity (or boundedness). Telic predicates, such as *Mary built a house* or *Ian ran to the store*, describe an event as having an inherent ending boundary, while atelic predicates, such as *Mary worked*, or *Ian ran*, describe an action without any reference to a specific ending. The telicity of a predicate is not only a semantic property, but is highly correlated with a range of syntactic behaviors, including argument structure, case-marking, auxiliary choice, and acceptability and/or interpretation of co-occurring adverbial phrases. When a concept plays a critical role in organizing grammars cross-linguistically, it is an excellent candidate for being a concept that organizes cognition more generally (Talmy, 2000; Lakusta & Wagner, 2013). And indeed, there is evidence that preverbal infants can track a range of event properties that are important for verb semantics (Hofer et al., 2005; Lakusta et al., 2007; Pruden et al., 2013, inter alia). Moreover, as children learn the specific elements of their language that signal telicity, they use those specific markers to guide their identification of individual event units, just as adults do (Wagner & Carey, 2003; Wagner, 2006). More generally, there is evidence that language labels facilitate event categorization among very young (3-month-old) infants (Ferry et al., 2010), which opens the possibility that learning event-related words may, in and of itself, be useful in the event segmentation process.

When adults are asked to segment a continuous stream of action into different events, it can be difficult to know precisely what cues they are using to do so, as there are strong correlations between all cues. For example, one event tested in a classic study of event segmentation (Zacks, Tversky, & Iyer, 2001) included making a bed. The identification of this event benefits from a conceptual understanding (it is easier if you are familiar with common housekeeping chores), an appreciation of the actor's intentions (which helps anchor the conceptual understanding), a common linguistic label ("making the bed"), statistical regularities within the sequence (the sub-parts of bed-making involve repeated actions over the same object while the end of the event likely involves the introduction of a new object), familiarity with the actions involved in the event (having the experience of making a bed yourself) and the kinematics of the biological action signals (which help signal the actor's intentions toward the bed). Adults could use any or all of these cues to understand the events going on around them.

By contrast, infants do not have equal access to all available event cues. As the literature cited above makes clear, there is good evidence that infants have some access to a wide range of cues, but (intuitively) the top-down cues involving a conceptual, intentional, and linguistic analysis of events are more dependent on one's life experiences. For an infant, the act of making a bed may be a rather exotic type of chore for which they have no label. Their ability to perceive the actor's intentions would rely primarily on the kinematics of the biological motion and statistical regularities of the objects involved.

We do know that these bottom-up perceptual cues can be an effective strategy for event segmentation, even for adults. For example, in Zacks et al. (2004), adults were able to consistently segment an abstract animation of dots undergoing randomly generated movement into units. But is still not well understood how even adults integrate these bottom-up cues with top-down cues. When participants in the study were given an intentional cover story for the random motion ("it shows a game"), their agreement about the event boundaries generally worsened. Zacks and colleagues suggest that taking an intentional stance may change how adults approach the task. Moreover, it's possible that intentionally created motion has somewhat different motion properties than random motion does, and therefore adults might have been facing a conflict between what the bottom-cues were signaling about event boundaries with what their intentional stance was used to identifying.

One solution to this problem is to find genuinely intentional events that are deeply unfamiliar to adults. Such events represent something much closer to the problem that infants face. The adults around babies are in fact engaged in intentional actions, but for purposes that the infant may not have any conceptual or linguistic access to. One example of an event of this sort was already discussed: Levine et al. (2017) showed adults ice-skating routines which are composed of units that adults have relatively little familiarity with. Bläsing (2015) used sequences of modern dance for the same reason. In both situations, novices were able to identify many of the same event units that experts identified. That is, even though the experts presumably were able to recruit a top-down conceptual understanding of the action sequence while the novices were dependent on a bottom-up perceptual analysis, they nevertheless were able to largely agree with one another. These results are encouraging if one is trying to imagine how infants break into the system of event understanding. The perceptual units that they can extract are in fact likely to be useful ones for creating a conceptual understanding. Moreover, we note that neither adults nor infants are wholly unfamiliar with dance. Kim & Schachner (2023) have shown that infants engage in dancing behaviors within the first year of life. More pointedly, Schachner & Carey (2013) showed that when adults were faced with actions that carried no other apparent purpose, they analyzed the actions as the core intention of the actor. Thus, adults perceive some motions as being done for their own sake, without any further goal.

However, in both studies that required participants to find specific units within unfamiliar action types, novice adults did not perform as the experts did. In Levine et al. (2015), novices agreed more with experts about the endings than the beginnings of events. And in Bläsing (2015) there were many points of divergence in segmentation going both directions—that is, times when the experts found dance units that the novices did not, and times when the novices found units the experts didn't recognize. More generally, the dance experts found fewer boundaries overall, suggesting that they were organizing the sequences into broader categories that were presumably conceptually meaningful. Moreover, Bläsing et al (2015) restricted their analyses to comparing only the number of segment boundaries identified between experts and non-experts, and in non-experts before and after learning. Their findings suggest that motor knowledge is one cue that affects event perception in modern dance but leaves open many questions about whether adults can normatively find intentional boundaries within novel dance sequences, and what other types of cues facilitate event segmentation of modern dance.

In the current study, we followed the lead of Bläsing (2015) and asked adults to segment modern dance into individual steps. Although human movements are expected to be familiar to our participants, and even though dance in a general sense is known to them, modern dance is less likely to be familiar. Moreover, unlike technical dance forms like ballet, which use a specific repertoire of codified steps (e.g., a *pirouette*), most modern dancers do not rely on a specific movement vocabulary. Instead, choreographers and dancers create sequences of new movement material when creating a novel work. Therefore, naïve observers cannot use prior knowledge of traditional or codified dance moves to identify the boundaries between steps in a modern dance sequence. Similarly, there are no pre-established linguistic labels for the steps in such a dance sequence.

In our first experiment, we modified the event segmentation task pioneered by Newtson (1973) and refined by Zacks (beginning with Zacks et al., 2001). In this general task, participants are shown a continuous stream of ongoing activity and asked to press a button whenever they perceive a boundary; that is, whenever they perceive that one meaningful unit has ended and another one has begun. For many events, adults show strong agreement about where these event boundaries occur, suggesting a common basis for analyzing actions and segmenting them into units. In the current study, we asked six professional modern dancers to choreograph and perform a dance sequence comprised of four unique steps. Dancers also performed each step in isolation to demonstrate their intended boundaries within the sequence. Participants were then asked to segment videos of the continuous dance sequences, and we compared their segmentations to the intentions of the dancers. We asked: Can adults find the units that the dancers intended? Do adults agree on where the boundaries are between steps? Our paradigm approximates the situation of infants, whose task also is to find the event boundaries that align with adult intentions.

In Experiment 2, we investigated more directly the relationship between linguistic conceptualizations and event segmentation for adults. In this experiment, adults were shown the same dance videos from the previous experiment, but instead of segmenting the dance in real-time, they were shown a static version of each dance in a series of still frames and were asked to identify units within the dance that they could linguistically label. The results from this experiment help us identify which parts of the dance sequences people can link to actions they can easily categorize with familiar language. Moreover, by comparing the results of the two experiments, we gain insight into the extent to which this level of linguistic categorization interacts with the ability to identify motion units in real-time.

2.0 Experiment 1: Online Event Segmentation

2.1 Method

The data presented in the study, stimuli, and analysis files are openly available on the Open Science Framework: https://osf.io/z4yu6/.

2.1.1 Participants

55 adult participants (31.5% female, mean age = 30.28 years, range = 18-56 years) participated in the current study. A slight majority (62.96%) identified as White. The remaining participants identified as African-American, (14.81%), Latino/Hispanic (9.26%), and Asian (7.41%); 5.55% identified as other. Overall, the participants were dance novices. Only 29.09% of adult participants (N = 16) had taken any kind of dance classes at any point in their life and only one (0.02%) was currently doing so regularly. More participants had watched dance performances (N = 40) but about half did so less than once per year (N = 19). Participants were recruited among the visitors to the Center for Science and Industry, a science museum in Columbus, Ohio. Verbal informed consent was obtained prior to participating in the experiment. All research and consent procedures were approved by The Ohio State University Social and Behavioral Sciences Institutional Review Board.

2.1.2 Stimuli

Stimuli were created by filming video clips of six professional modern dancers from Syren, a modern dance company based in New York City. Dancers were filmed at the Ohio State University's STEAM factory. In addition to providing dance stimuli, dancers also participated in individual interviews about their backgrounds and dance philosophies. Dancers were asked to choreograph and perform a short dance containing four unique steps. After recording each dancer, we then asked them to demonstrate the four steps by pausing at the boundary between each step (see Figure 1). The videos of the continuous dances were used as stimuli in the experimental procedure. The recordings of the individual steps were used to define the 'ground truth' of the step boundaries during data analysis. All six videos were edited to be approximately 6.5 seconds in length using Adobe Premiere software. Previous literature has shown that higher level information (e.g., from emotions; Christensen et al., 2016) and statistical information (Opacic et al., 2009) can modulate processing of events lasting just a few seconds, suggesting that our dance stimuli should be of sufficient duration to elicit both bottom-up and top-down processing.

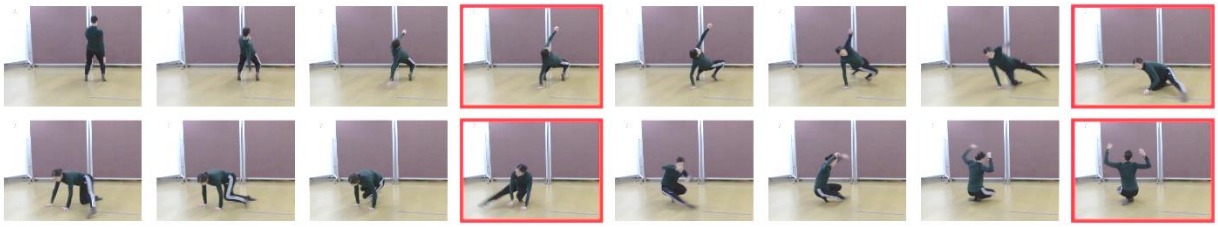

Figure 1. Example frames from one dance video, with the frame depicting the end of each step outlined in red. Dance videos varied naturally in the timing of the four steps.

2.1.3 Procedure

Participants were seated in front of a computer monitor and keyboard. Stimuli were presented using E-prime software. Participants were asked to segment each dance sequence by pressing a key whenever they perceived a dance step boundary (Zacks et al., 2009). The six videos were presented in randomized order, and each dance video was presented twice consecutively, generating two sets of segmentation data for each video. Participants completed two practice trials before beginning the task. The practice video was of another modern dance sequence that had been recorded for a different purpose. Keyboard responses were recorded with E-prime. After the task ended, participants were asked open-ended questions about their dance experience and history (see Supplementary Table S1).

2.2 Data Analysis and Results

2.2.1 Segmentation ability

Segmentation ability was defined as a participant's ability to identify a step according to dancers' intended boundaries. First, we divided each dance video into 12 even temporal bins, which resulted in bins of approximately 540ms (see Supplementary Table S2 for exact bin sizes for each video). Next, we calculated the proportion of participants who pressed a button within each bin (Sargent et al., 2013). For the ground truth, a bin was assigned a 1 if it contained a true boundary, and a 0 if it did not. Following the procedure of Zacks et al. (2009), we fitted cross-correlation sequences between participant responses and the dancers' 'ground truth'. This was done to account for the lag between the moment an observer perceives a step boundary and the moment they execute a button press, which is typically shifted in time and can occur before or after the true boundary (Zacks et al, 2009). Cross-correlation sequences were calculated using lags from -11 to 11 bins. Participant responses were then shifted in time using the lag with the largest correlation, to maximize the correlation between responses and the ground truth. Correlations and optimal lag times for each dance video are reported in Supplementary Table S3. 83% of the lag times were zero, meaning that the data did not need to be shifted.

Figure 2 depicts histograms of the number of participant key presses for all six dance videos and gives a holistic sense of how participants performed on this task. As can be observed in the figure, all steps were identified by at least some participants, but certain steps were identified more frequently than others. For example, participants appear to have collectively identified the fourth step of dances three and six, as indicated by the higher frequency of key presses during the bin containing those step boundaries. As revealed in Figure 2, nine of the 24 'correct' steps were identified at high rates: more than 1 SD above the mean number of key presses across all bins (mean = 11.40, *SD* = 6.91). Two bins, corresponding to the final dance step of Dance 3 and 6, were identified at very high rates (> 2SDs above the mean), indicating that perception of these event boundaries was particularly strong. Out of the remaining 48 'false alarm' bins, which did not correspond to step boundaries intended by the dancers, two bins were identified more frequently than 1 SD of the mean, during Dance 1 and Dance 4. There were no ‘correct’ dance steps that were identified at a rate *lower* than 1 SD below the mean.

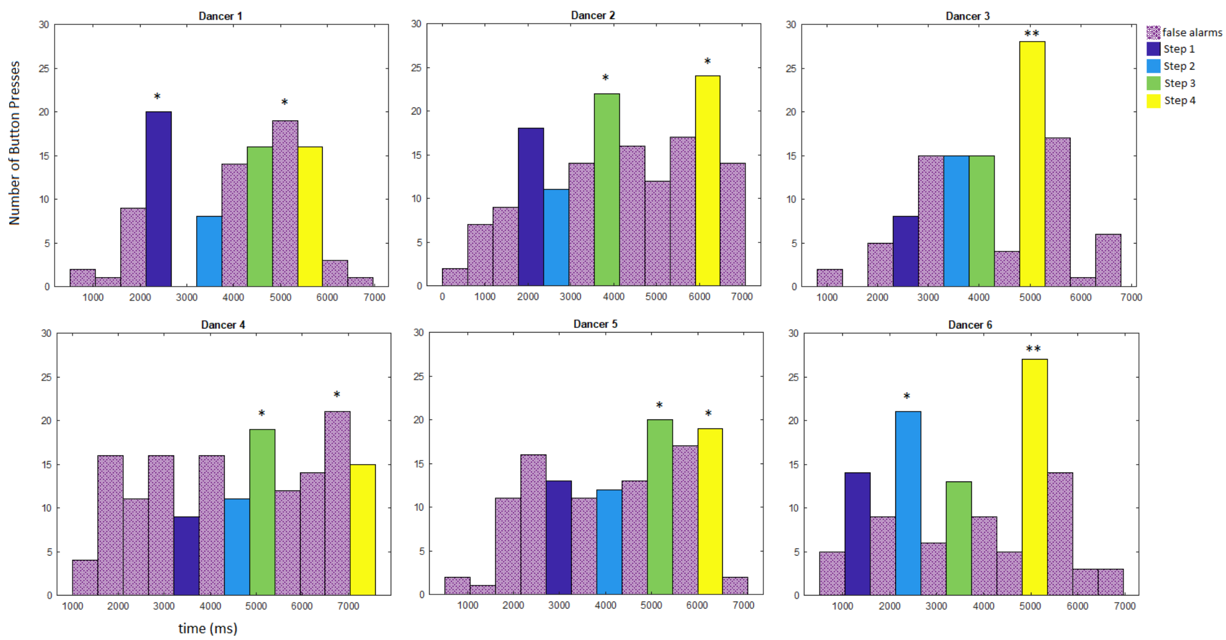

Figure 2. Response distribution across all participants, divided into 520ms bins. Colored bars indicate bins containing a ‘true’ step according to the dancers’ intentions. *Denotes bins that were identified at high rates (> 1 *SD* above the mean frequency across all bins) ** = 2 *SD*.

After optimally shifting participant responses, performance was evaluated via (*d*’), derived from signal detection theory (Macmillan, 1993), a measure which accounts for correct responses (‘hits’) and incorrect responses (‘false alarms’)[1]. The *d’* score can be either positive or negative, and a more positive *d’* reflects a more sensitive response. We also calculated a ‘hit rate’ for each dance step by counting the total number of hits per step out of the total number of

[1] *D*’ was calculated as follows (Macmillan, 1993):
$d' = \Phi^{-1}(H) - \Phi^{-1}(F)$ where Φ^{-1} represents a z-score function, H represents the hit rate, and F represents the false alarm rate. In other words, d’ is calculated by subtracting the false alarm z-score from the hit rate z-score.

participants. A correlation analysis revealed that the number of total hits during the first and second presentations of each dance video were highly correlated ($r = 0.87$, $p < .0001$). Therefore, for simplicity, we only considered data from the second round in subsequent analyses.

Participants' mean sensitivity (*d*') was 0.50 (*SD* = .42). A one-sample t-test revealed that sensitivity differed significantly from zero, $t_{54} = 8.90$, $p < .0001$, indicating that participants did discriminate step boundaries at above-chance levels.

2.2.2 Analysis of Kinematic Movement features

After determining *whether* participants could segment the dance videos, the next question we asked was *which features* of the dance steps enabled them to do so? Our coding categories were inspired by the classic Laban analysis of dance movements, a widely used system for the description of human movement (Laban & Ullmann, 1975). To examine this question, two independent research assistants coded each dance step for the presence of twelve specific kinematic features, such as 'spin' or 'arm movement' (see Table 1 for a complete list with definitions). Any disagreements between the two coders were resolved via discussion between the authors. Each dance step was then assigned a 1 if it contained a particular kinematic feature and a 0 if it did not.

Table 1. *Definitions of movement features and number of steps coded as displaying each feature.*

	Movement feature	Definition	# Steps
1.	Position switch	The dancer changed from upright to horizontal position (or vice versa).	9
2.	Jump	The dancer left the floor.	3
3.	Spin/roll	A rotation of the body.	5
4.	Torso movement	A movement of the torso.	18
5.	Arm movement	Movement of the arm or hand that is independent from the torso, like raising the arm.	18
6.	Head movement	Movements of the head that were independent from the torso, like turning the head to the side.	9
7.	Horizontal space	The dancer moved to a new space along the horizontal dimension.	13
8.	Vertical space	The dancer moved to a new space along the vertical dimension.	14
9.	Break	A spatiotemporal pause at the end of the step.	8
10.	Gravity	The step contained a movement that was determined by gravity, like a landing or a fall.	6
11.	Flow	There was a smooth, continuous transition at end of the step into the next one.	16
12.	Coda	There was a flourish at the end of the step.	5

To examine the contributions of each movement feature to participant responses, we created a categorical, binary classification for every participant and each dance step, assigning a 1 if the participant had a ‘hit’ at the bin associated with the step’s boundary point and a 0 if they did not. Using this classification as the dependent variable and the movement features identified for each step as categorical predictors, we fitted a mixed effects logistic regression model. This analysis was implemented in R using the R package *lme4* (Bates et al., 2015). The model was fitted with all twelve movement features as categorical fixed effects (present = 1, absent = 0), participant ID as a random effect, and participant response (0 = miss, 1 = hit) as the dependent variable. One feature—*flow*—was dropped from the model because it was not linearly independent from the other movement features. Table 2 summarizes results from the regression model.

Table 2 Mixed effects logistic regression predicting *hits* from movement features.

Fixed effects	Estimates (β)	95% CI	p
Intercept	1.073	[0.60 1.55]	<0.001 ***
Position switch	0.033	[-0.39 0.45]	0.876
Jump	0.640	[0.07 1.21]	0.028 *
Spin	0.392	[-0.06 0.84]	0.089 #
Torso movement	-0.205	[-0.73 0.32]	0.443
Arm movement	-0.034	[-0.43 0.37]	0.868
Head movement	-0.230	[-0.62 0.16]	0.252
Horizontal change	0.251	[-0.15 0.65]	0.217
Vertical change	0.003	[-0.39 0.39]	0.989
Break	0.140	[-0.30 0.58]	0.532
Gravity	-0.578	[-1.19 0.03]	0.063 #
Coda	-0.226	[-0.86 0.41]	0.485

#p<0.1, *p<.05, **p<.01,***p<.001

The parameter estimates (β) can be interpreted in terms of the expected log odds of a participant identifying the boundary of a step–that is, getting a hit–given the presence of a specific predictor (i.e., movement features). For example, the presence of a *jump* within a dance step was associated with a 0.64 unit increase in the expected log odds of a hit, which was a statistically significant increase (p = .028). The only other feature which came close to statistically improving segmentation (with a marginal p-value of .089) was *spin,* which increased the expected log odds of a hit by .392. On the other hand, a *gravity* was associated with a decrease in the expected log odds of getting a hit—in other words, participants were *less* likely to identify step boundaries for steps that contain these movement features. None of these effects are particularly strong, and they only partially account for the pattern showed in Figure 2. For example, the fourth step of Dancer 6 was well identified and, kinematically, it contained a spinning motion which is consistent with statistical results. But spinning was also a feature of the first step for Dancer 3, which participants found at fairly average rates. Moreover, some well-identified steps, such as the fourth step for Dancer 3, contained neither a spin nor a jump but involved a distinctive kicking motion.

2.2.3 Segmentation agreement

Our final question was whether participants tended to agree with one another on when step boundaries occurred, regardless of whether they agreed on the ground-truth provided by the dancers. Following the procedure of Sargent et al. (2013), agreement was defined as the correlation between an individual's responses and the group. To determine the group norm, we calculated the proportion of participants who identified a boundary within each bin. For each participant, we converted their responses for each dance video into a binary vector in which each bin was assigned a 1 if it contained a button press and a 0 if not. We then calculated the correlation between that participant's segmentation and the group norm. Correlations were rescaled to control for individual differences in the number of boundaries identified (following the procedure of Kurby & Zacks, 2011[2]).

Mean scaled segmentation agreement was 0.72 (*SD* = 0.22). In past studies, mean segmentation ability ranged from 0.58-0.59 in Sargeant et al. (2013); 0.50-0.67 in Kurby & Zacks (2011), and 0.39-0.43 in Zacks et al., (2006). This finding reveals that segmentation agreement in the current study was in fact higher than in other event segmentation experiments. Therefore, it can be said that participants generally agreed with one another about where the dance step boundaries occurred.

2.3 Discussion

In sum, adults agreed with each other about the location of step boundaries at rates that were similar to previous studies, but nevertheless had an average sensitivity score of only 0.5 for

[2]For each participant, segmentation agreement scores were rescaled to 0–1 agreement scores by first computing the correlation between that individual's binned event boundaries and the distribution for the group, *r*. Next, the minimum and maximum correlations possible were calculated, given the overall number of boundaries the individual had identified (r_{min} and r_{max}). Finally, the following formula was implemented:

$$SegmentationAgreement = r - r_{min} / (r_{max} - r_{min})$$

the boundaries that were intended by the dancers. This task, therefore, is a challenging one. Still, the dancers' steps did appear to be the dominant guide to adults' perception of event boundaries. Only two bins received high numbers of key presses that did *not* correspond to the dancers' intended boundaries, and we note that both bins occurred immediately prior to the final step in their respective dance. Thus, it is possible that responses in these bins represent anticipatory key presses of the final step in the sequence, which should in principle be easy to detect as it represents the end of the movement sequence.

Findings from the multiple logistic regression model revealed that the presence of only two kinematic features—jump and spin—significantly increased the odds that adults would identify a step boundary. These two features are also the only two types of movement that are entirely captured by common English words. This finding raises the possibility that *jumps* and *spins* were easier to quickly label with a word, unlike, for instance, the motion captured by the feature of *horizontal arm movement*. Experiment 2 directly investigates the kinds of labels people apply to these dance sequences.

3.0 Experiment 2: Event Labeling

In Experiment 1, participants were not asked to provide labels for the dance steps, but previous research has found that linguistic labeling can change how an event is remembered and analyzed (Papafragou et al., 2008; Papafragou & Selimis, 2010). Moreover, as noted previously, language provides rich top-down cues about which parts of an event are important, particularly with respect to whether the ending of an event is important (Wagner & Carey, 2003; Wagner, 2006). Given the real-time nature of the segmentation task in Experiment 1, it is unlikely that participants were carefully considering how to label each step. However, it is quite possible that steps which were readily perceived as common actions automatically invoked common verbal descriptions. Preliminary support for this idea comes from the fact that steps containing motions that could be labeled by the common verbs *spin* and *jump* were segmented more successfully. If some steps were indeed more readily labeled by participants, then these labels may have contributed to their segmentation successes.

In Experiment 2, we explicitly asked participants to provide labels and to link those labels to perceived steps. Of particular interest was whether the steps differed in the extent to which participants perceived them as label-able; and, for steps which did regularly receive labels, whether there were differences in the linguistic properties of those labels. If we assume that any linguistic influence on segmentation would have occurred relatively quickly, then steps which people associated with easily accessed labels (such as commonly used verbs) might be associated with improved segmentation performance. Moreover, as segmentation depends on finding event endings, then steps which people labeled with linguistic phrases that explicitly targeted endings (i.e., telic predicates) might also be associated with improved segmentation performance.

3.1 Method

3.1.1 Participants

50 adult participants (24 female, mean age = 28.7 years, range = 18-56 years) participated in Experiment 2. Half (50%) of the participants identified as female, and the majority (86%) identified as White. The remaining participants identified as Latino/Hispanic (8%), and Asian (4%); 2% declined to report. As with the previous study, these participants were also dance novices. Only 24% of participants (N = 12) had taken any dance classes at any point in their life and very few were currently doing so (N = 3). More participants had watched dance performances (N = 32) but most did so less than once per year (N = 24). Participants were recruited via Prolific, an online research platform (prolific.co). Consent was obtained prior to

participating in the experiment. Participants were compensated for their time at $12 per hour. All procedures were approved by The Ohio State University Social and Behavioral Sciences Institutional Review Board.

3.1.2 Stimuli

The dance videos from Experiment 1 were used as stimuli and presented to participants using Qualtrics. For each video, every 20th frame was extracted and aligned in a row to create a still frame representation of each dance video, similar to the array of still images depicted in Figure 1. Each dance was represented with a sequence of between 13–15 still frames.

3.1.3 Procedure

After providing consent, participants watched each dance video and were then asked whether they noticed any dance step that could be labelled with a common English word or phrase. If they answered yes, they were asked to enter the label and to select the corresponding frames of the dance from the set of still frames. Participants were then asked if they perceived any other steps that could be labeled. If they again answered yes, they were again asked to provide the label and mark the corresponding dance frames. Participants were allowed to provide up to four labels per dance before being moved to the next video. The six videos were presented in the same order to all participants, and participants could re-watch each video as many times as they desired prior to answering the questions. After participants had watched all six dances, they were asked the same demographic questions, including questions about their dance experience and history, as in Experiment 1 (see Supplementary Table S1).

3.1.4 Coding of Labels

Each label was coded for its general semantic domain. The domain codes were the following: COMMON vocabulary, that is, words that are high frequency and can be used in a variety of situations (*kick, jump*); SPORTS/EXERCISE vocabulary, that is, words that are strongly associated with particular sports or with calisthenics (*downward dog, push-up)* as well as the few instances of explicitly dance-oriented terms (*pirouette*)*;* DESCRIPTIVE phrases focusing primarily on the physical properties of the motions (*floor leg fan);* and POETIC items which described the step in a more metaphoric manner (*monk, snake, mayday).* Only two items did not fit into one of these categories and were omitted from analysis.

All labels were also coded for their telicity value. The telicity coding, however, was done in a generous spirit. As can be inferred from the semantic domain codes above, many of the labels were not verb phrases of any sort. The standard linguistic tests for telicity (Dowty, 1979) apply only to verb phrases, making their application questionable at best with much of these data. Thus, each label was primarily classified conceptually (did it describe an event with a specific ending or not?). This classification was verified to the extent possible by the application of the *in/for a minute* test: for atelic predicates, *for a minute* is the most natural way to describe the duration of the event, while for telic predicates, *In a minute* sounds more natural. Thus, labels which described events as actions with no reference to their endpoints (*slide, swim)* were coded as ATELIC. Labels which identified a specific ending (*make the wheel)* were coded as TELIC, as were labels which described punctual actions (*push-up, cartwheel, backflip)*. A few labels could not even generously be interpreted as verb phrases (*swoosh, floor listener*), and were classified as Uncodeable.

3.2 Data Analysis and Results

Eight participants provided no labels, but the remaining 42 provided a total of 198 labels, for an average of 4.7 labels per person. Participants were allowed to freely choose the number of frames they wished as the referent for their label. These frame sets were inspected and compared to the frames that corresponded to the dancers' ground-truth for the steps to determine which step was most likely being labeled. Frames sets which exactly corresponded to one of the dancers' boundaries or which were a proper subset within those boundaries were coded as representing that step. If participants chose frame sets that overlapped equally with multiple steps, the label itself was the deciding factor: whichever step it described most clearly was assumed to be the intended referent. This method of assigning labels to steps already builds in a certain level of accuracy relative to the dancers' ground truth because a label had to include at least one frame from the dancer's truth to be linked to it[3]. However, there was variability in how well the labels reflected the dancers' units. Therefore, we also calculated a *d'* score to determine how well the labels matched the dancer's intended steps, following the same procedure of Experiment 1. For this analysis, a 'hit' was defined as a frame that the fell within the boundaries identified by the dancer, including the end-point; a 'false alarm' was scored if the participant selected a frame outside of that set. Across all the labels provided, the mean *d'* score was 1.57 (SD = .56). This sensitivity significantly exceeded a chance level of zero, $t_{197} = 39.56$, $p < .001$, CI = [1.49 1.65]. Thus, overall, the labels were reasonably reflective of the intentions of the dancer.

Figure 3 shows all the labels distributed across the steps from the six dancers. Every step was labeled at least once, although the number of labels a step received varied widely, from 1 to 29 labels. The mean number of labels per step was 8.25 (SD = 6.4), the median number of labels per step was 6, and the modal number of labels per step was also 6. We used the median value to

---

[3] To ensure that we were not artificially inflating our significance levels with this method of anchoring labels to steps we also conducted an additional analysis in which each label was randomly assigned to one of four possible steps per dance. The d' for this random set was 0.38 and distributed roughly evenly around 0. Moreover, it was significantly lower than the d' noted in the main text ($t(197) = -16.11$, $p < .0001$).

classify a step as being relatively easy or hard to label. Steps which received fewer labels than the median (< 6) were classified as relatively difficult to label (*N* = 9). The remaining 15 steps each received 6 or more labels and were classified as relatively easy to label.

To compare this labeling classification to the segmentation data, we took the d' scores for each step in Experiment 1 and transformed them into z-scores to generate an ordering of the steps that captured the ease with which they were segmented: positive values indicated that the step was correctly identified more than the mean of all the steps and negative values indicated that the step was identified less frequently than the mean. We then calculated a t-test using the labeling difficulty (easy vs difficult) as the independent variable and the segmentation z-score as the dependent measure. (Note that because different participants were in each study, this analysis is an item analysis over the 24 steps.) The results showed that there was a significant effect of label-ability (t (23) = 2.2, p =.04). The average segmentation z-score for the easy-to-label steps was .32 (*SD* = 1.04) which was significantly higher than the average segmentation z-score for the difficult-to-label steps (-.54, *SD* = .67). In sum, steps which adults found easier to label in Experiment 2 were more likely to be the steps that adults succeeded in segmenting from the continuous stream in Experiment 1 (Figure 4).

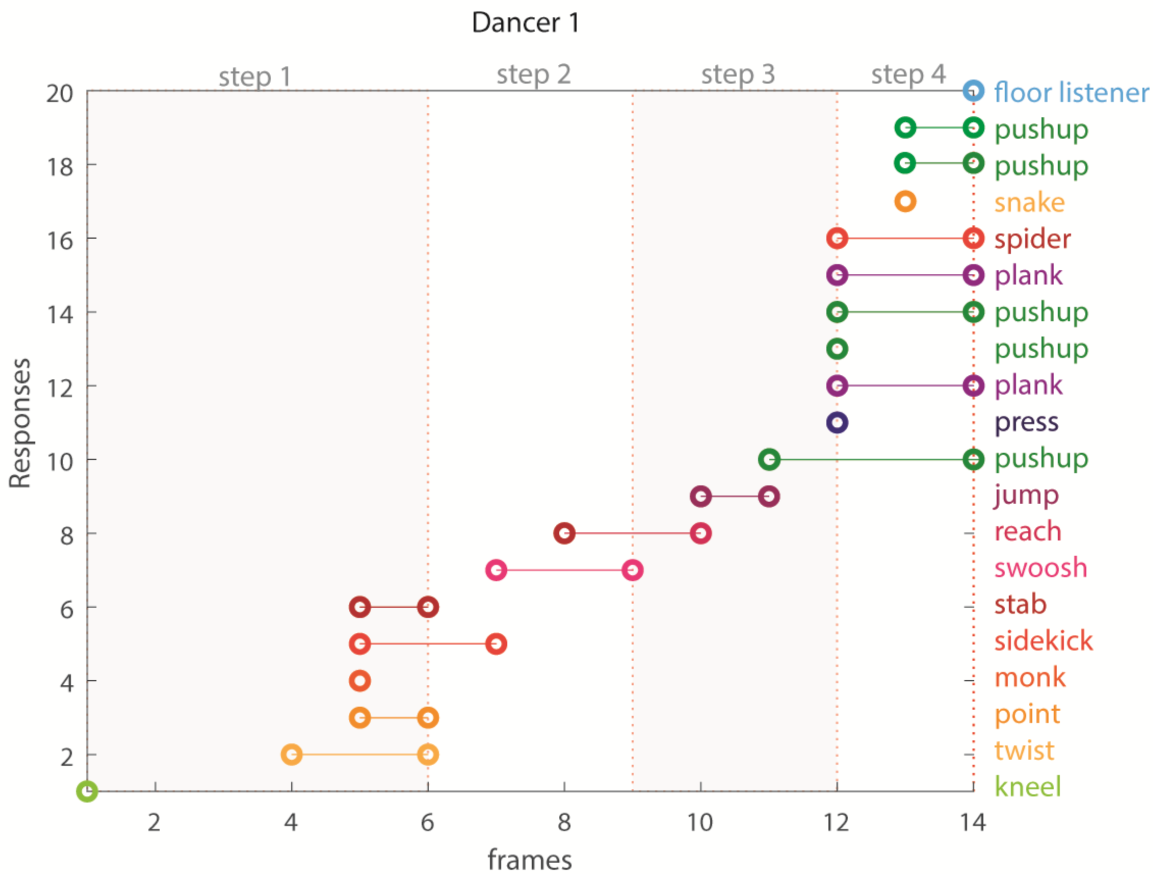

Figure 3: Responses from Experiment 2 for a representative dancer. Each row depicts the frames selected for each response, with the corresponding label printed to the right. Colors indicate unique labels. Dotted lines indicate the step boundaries corresponding to the dancer's intentions. Graphs for dancers 2-6 can be viewed on https://osf.io/z4yu6/.

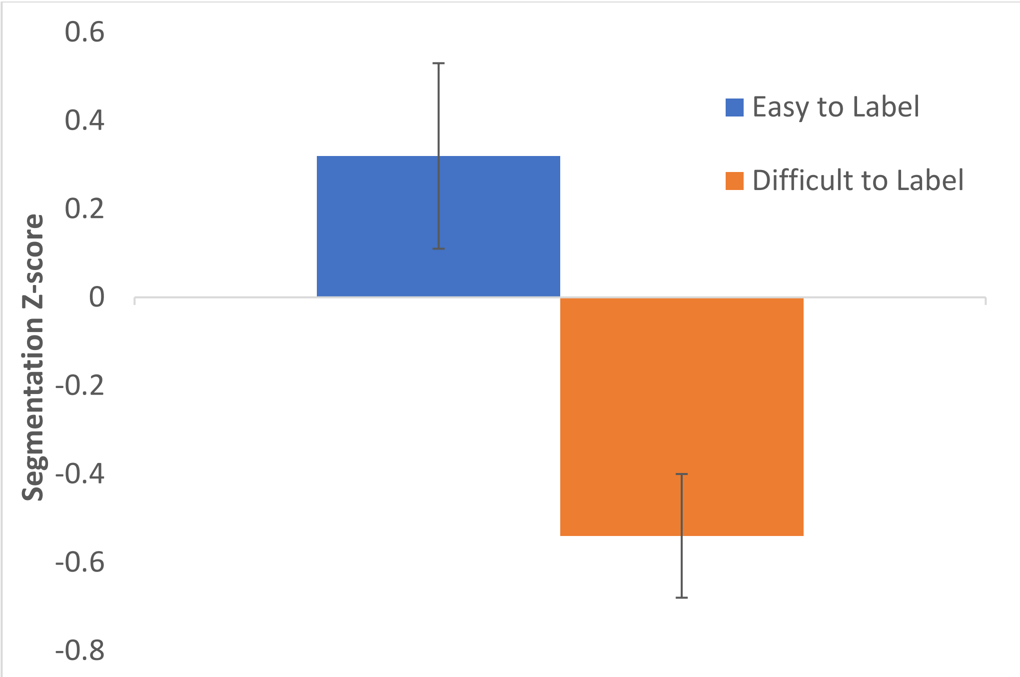

Figure 4: The relationship between how easy or difficult a step was to label in Experiment 2 and how successfully it was segmented relative to other steps in Experiment 1.

Our analysis of the telicity of the labels found a roughly even split between labels that were plausible Atelic (39.4%) and labels that were plausibly Telic (47.0%). The remaining 13.6% of labels were classified as uncodeable. We were able to verify the validity of our telicity coding by asking if participants were more likely to include the frame corresponding to the end-point of the step in their reference set when they provided a telic label (which linguistically specified an ending) compared to an atelic label (which did not). Inspection of the means showed that 65% of Telic labels contained the end-point frame while only 47% of Atelic labels did so. This difference was significant: ($F(1, 170) = 5.13$, $p = .025$). However, the telicity classification did not predict how accurately participants identified step boundaries within the still-frames ($F(1, 170) = 2.3$, $p = .13$). Unfortunately, as over a third of the steps did not receive even five labels, it was not statistically feasible to do a cross-experiment item analysis to test whether the telicity classifications would predict on-line segmentation.

Finally, we examined the specific nature of the labels being used. Overall, labels were primarily coded as either Sports/Exercise terms (47% of labels) or Common verbs (32.3% of labels). Only 13.1% were coded as being Poetic and 6.6% of labels were coded as Descriptive phrases. Overall, there was not a strong consensus among participants about the type of labels to use. A total of 92 distinct labels were provided and each label was given an average of 2.2 times, with the modal number of instances per label being 1. Among the steps that received at least 5 labels, only twelve (half of the full set) were labeled by at least 50% of participants with the same general type of term: nine with Sports/Exercise terms and three with Common verbs. One consistent pattern we observed was that most labels (all the Common verbs and most of the Sports/Exercise terms) were everyday, high frequency words. Indeed, five words accounted for 40% of all the labels provided: *cartwheel, pushup, roll*, *slide* and *spin.* However, with the exception of *cartwheel,* all of these labels were used to label multiple steps: They appear to be somewhat general in their application. The label *cartwheel,* however, was used exclusively with Dancer 2's first step. Inspection of the segmentation data shows that this step was identified reasonably often, but not as frequently as other steps (it received fewer than 1 SD more than the mean number of key presses). Nevertheless, the fact that people rely on familiar labels in the off-line version of this task supports the idea that only such items would likely be accessible in the on-line segmentation task.

3.3 Discussion

Experiment Two removed all time pressures from participants and asked them to consciously adopt a top-down linguistic approach to their analysis of the dance. The results showed that participants could provide labels that corresponded with reasonable accuracy to the dancers' intended steps. Moreover, steps which received relatively more labels were more likely to have been segmented in the online task of Experiment 1. One potential interpretation of this result is that language has a facilitative effect on segmentation: being able to come up with a label for what you see makes it more likely that you will correctly see the intended boundaries of the event. However, these data are purely correlational, so it is equally possible that seeing the intended boundaries facilitates one's ability to produce a label for it. Or, it may be that there is a third feature that correlates both with label-ability and segmentation: in the general discussion, we will suggest that familiarity could be that third feature.

4.0 General Discussion

How do we perceive structure within continuous human movement that is highly unfamiliar to us? Infants are regularly faced with this task and succeed at finding meaningful actions that correspond to what the people around them are doing. A variety of cues can be used to help find the boundaries of individual events, ranging from an appreciation of human kinematics to recruiting linguistic descriptions of the actions. To investigate how different cues might be exploited, we conducted two experiments that placed adults in a situation analogous to that of an infant: adults were presented with fluid sequences of modern dance and asked to identify the boundaries between dance steps (Experiment 1) and to provide labels for any perceived dance steps that could receive them (Experiment 2). Findings demonstrated that segmenting these unfamiliar movement sequences is challenging and that adults benefit from being able to access top-down information, particularly linguistic information.

Experiment 1 implemented a traditional event segmentation task (Sargent et al., 2013, Zacks et al., 2009) and evaluated participants' success at identifying the internal dance steps relative to the intentions of the dancers. Overall, participants (as measured by a *d*' score), showed that they were reasonably sensitive to where the step boundaries occurred in the sequence.

Segmentation agreement, as traditionally measured, also indicated that participants did agree with one another at rates that have been interpreted in prior segmentation studies as demonstrating high levels of agreement.

These findings raise interesting questions about the nature of event segmentation. One major difference between the current study and prior event segmentation studies is that our goal in Experiment 1 was to compare participants' segmentation with an objective 'ground truth'; that is, with the intentions of the dancers who created the movement sequences. Most prior studies have instead used segmentation ability as another dependent variable with which to predict other measures of event perception or memory (e.g., Sargent et al., 2013; Swallow, Zacks, & Abrams, 2009). Our analysis revealed that participants did segment some events that corresponded to the dancers' intended boundaries, although they also segmented two other events that did not coincide with the dancers' 'ground truth' (cf. the particularly high bars in Figure 2). This suggests that participants perceived event boundaries that were not created intentionally by the dancers, and as a group they also missed other intended boundaries. These findings suggest that, despite reasonable levels of agreement among one another, perceiving event boundaries that coincide with the intentions of others is in fact a challenging task that adults were only moderately successful at. We do feel compelled to note that this finding does not speak to the aesthetics of the dances. Anecdotally, people do seem to enjoy watching these short dances, suggesting that appreciating the dancers' intended steps is not a necessary element of dance appreciation.

To understand what made these sequences difficult to segment in terms of their intended boundaries, we must consider how different kinds of cues might be used to solve the problem. One low level perceptual cue that would have been almost useless in this task was spatiotemporal boundaries between the steps, as the dance sequences were very fluid. Indeed, the dancers themselves described fluidity and connectedness as being something they aspire to within their dancing. In the words of the dancers themselves, they wanted to create a “seamless” “flow” with a “through-line” that was “as smooth as possible.” While large pauses between action steps is unusual even for everyday action sequences (it would be quite odd for someone to pause in stillness between pouring the milk on their cereal and returning the milk to the refrigerator), the fact that the dancers were purposefully trying to blur the lines between their steps may have removed even subtle breaks between the steps. Since we did not conduct a complete motion analysis of our sequences, we cannot rule out the possibility that such subtle breaks did exist. However, given our results, such breaks were apparently not sufficient to make the task an easy one. Moreover, this approach of the dancers contrasts sharply with the figure-skating routines in Levine et al. (2017). The individual “steps” in those routines are regularly separated by neutral bouts of skating which may have served as a motion baseline, allowing the participants in that study to extract the target steps more easily from the background.

Another bottom-up cue that should in principle have been more useful was the kinematics of biological motion. From the perspective of Event Segmentation Theory (Zacks et al., 2007), our participants’ implicit understanding of how the human body moves and the constraints of physics should have made the motion sequences predictable enough to allow participants to anticipate when motion steps would end. That is, no matter how fluid the dancers were trying to be, they were still subject to the ways the human body can move and the fact that gravity would return them to the floor. Our participants’ weak success suggests that they were not able to leverage this cue effectively. Moreover, our analysis on how different kinematic properties predicted segmentation success showed that even steps containing dramatic kinematic cues–e.g., major shifts between a vertical and horizontal orientation, being brought back to the floor by gravity–were not more successfully parsed out.

One possible reason that the participants were not able to make strong use of these kinematic cues may have to do with how the perceiver's own motor experiences may influence their perception of motion. An extensive body of research has identified a tight coupling between action execution and action observation at the neurological level (Kilner et al., 2009; Sebanz et al., 2003). For example, adults show preferential activity in the motor cortex for actions they are experts in (Calvo-Merino et al., 2005) and even infants show increased motor activity while observing actions that they have recently acquired in their motor repertoire (Gerson et al., 2015). Our dance sequences were created by full-time professional dancers who each had many years of dance training. By contrast, our participants were dance novices who had no (or extremely little) training and did not engage in dancing as part of their regular life. Thus, our participants' lack of personal motor experience with the motions of the dancers may have reduced their ability to exploit the kinematic motion cues in the dances. Further research investigating the segmentation skills among modern dancers would clarify this point.

If this line of reasoning is correct, it suggests that additional motor experience would facilitate segmentation of dance steps. Interestingly, when the Syren dancers were asked informally how they would teach a novice their steps, one common strategy mentioned was using metaphors to make it easier for students to understand what to do with their limbs and body position. For example, one said she might describe a dance move as "bend down and swat a fly". These teaching approaches suggest that dancers are aware that internal motor simulation and perceptual resonance facilitates learning: mapping new movements onto other movements already within their motor repertoire is a powerful way to teach novice dancers (Fisher, 2017).

A third cue that participants might use involves an appreciation of the dancer's intentions. Given that the dances were enacted in a purposeful manner, their intentions would presumably be useful information. However, the nature of the dance motions may have made the identification of intentions particularly difficult. There is some evidence that when adults are faced with a seemingly purposeless action (a "mysterious" action), they are able to conceptualize the actors' intention as being to produce the action itself (Schacher & Carey, 2013). But as the authors note, these kinds of intentions are relevant in only a limited number of situations, such as exercising, various rituals, and of course, dancing. More commonly, everyday intentional actions center do in fact center on objects: things we move, eat, change, prepare, and otherwise interact with. For example, in a classic study on infant event segmentation (Baldwin et al., 2001), the key events that infants successfully segmented involved picking a towel up from the floor and placing a tub of ice cream in a refrigerator. Similarly, in Woodward, (1998), infants preferentially analyzed a simple event as reaching for an object as opposed to simply moving one's arm in a particular direction. Moreover, even adult studies of event segmentation generally feature everyday activities that center around objects, such as doing laundry, building a Lego house, making breakfast, or gardening (e.g., Sargeant et al., 2013, Zacks et al., 2009). Thus, while adults are capable of interpreting dance sequences as intentional overall, without object cues to anchor their analysis, adults may have struggled to determine specific intentions and thus the specific boundaries within the sequence.

The final cue to consider is language. The results from the kinematic analysis found that just two types of kinematic changes were linked to improved event segmentation: jumping (significantly so) and spinning (marginally so). We found it notable that these two motions can be readily labelled with common verbs, namely "jump" and "spin". By contrast, words describing various arm motions and torso shifts are not as readily coded into simple language. It is possible, therefore, that what helped participants identify steps was their ability link them to a label (and see also Papafragou & Selimis, 2010 and Papafragou et al., 2008 for further discussion about how linguistic encoding could influence event perception). Experiment 2 explored this possibility in detail by asking people to find steps in the dance sequence that they could provide labels for. And indeed, we found a relationship between how easily a step could be labeled and its ability to be segmented according to the dancers' intentions. Steps which were labeled by more people in Experiment 2 were segmented more often in Experiment 1.

If language is facilitating segmentation, however, it appears to do so in a general way. There was not much consistency in the specific labels that people provided, and the properties of the labels–their general domain of reference, and whether they linguistically targeted an end-point–were also too inconsistent across participants to predict segmentation success. These results suggest that language is not helping by providing any kind of fine-grained guidance to the dance. And perhaps the relatively short duration of the clips themselves limited participants' ability to use language. Alternatively, these results may suggest that it is not language which facilitates segmentation, but the identification of event units that inspires people to search for a linguistic term of some kind. There is evidence that children's word learning benefits from discretely segmented word forms: when words are presented to them in isolation, they find it easier to learn their meanings (Brent & Siskind, 2001). Moreover, Tomasello & Kruger (1992) found that children learned verbs better when they were presented outside the boundaries of the event they labeled. Allowing children to analyze the event stream without distraction helped them link it to a novel verb.

In addition, it is also possible that there is a third factor at play here. After all, the two kinematic features linked to improved segmentation, jumping and spinning, are not only easily labeled motions but are also motions that are highly familiar to our participants. Moreover, that familiarity likely encompasses not only experiences of observing these actions but also of personally engaging in them. Rather than language facilitating segmentation or vice versa, it may be that visual and/or motor familiarity with specific actions facilitates both segmentation and labeling. Teasing apart the precise nature of the relationship of language and segmentation and the potential contribution of familiarity would require a more structured set of materials as well as a manipulation involving variable exposure to the different steps.

Nevertheless, the fact that language (or arguably, familiarity) was the only cue that even has the potential to provide real traction for participants in this task illustrates just how hard it is to segment completely unfamiliar motion sequences. Even for adults with ample experience with a wide range of motion types and with years of event segmentation to draw on, a truly unfamiliar action sequence does not easily afford individual units. These results might give us pause when thinking about how confident we are when interpreting activities from contexts or cultures we are highly unfamiliar with. They should also impress us with the abilities of infants who consistently solve the event segmentation problem.

Acknowledgements

This project was funded by an Ohio State University Social and Behavioral Sciences grant to LW and NIDCD grant number F32DC017076 to CM. Student support was provided by an NSF REU Site grant (#SMA-1757020**).** We thank Savannah Vasquez, Katriese DeLeon, Victoria Greyson, and Hannah Kosstrin for help with data collection and coding. Thanks to Jeffrey Zacks for sharing his E-prime experiment. We are grateful to the Language Sciences Research Lab at the Center for Science and Industry (COSI) for providing laboratory space. Finally, we are grateful to the Syren dance company for generously providing the dance materials.

References

Adolph, K. E., & Hoch, J. E. (2019). Motor development: Embodied, embedded, enculturated, and enabling. *Annual review of psychology*, *70*, 141-164.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407. https://doi.org/10.1016/j.cognition.2007.07.005

Baldwin, D., Baird, J., Saylor, M., & Clark, M. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717. https://doi.org/10.1111/1467-8624.00310

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bläsing, B. E. (2015). Segmentation of dance movement: Effects of expertise, visual familiarity, motor experience and music. *Frontiers in Psychology*, *5*(OCT).

https://doi.org/10.3389/fpsyg.2014.01500

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33-B44.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex*, *15*(8), 1243–1249. https://doi.org/10.1093/cercor/bhi007

Christensen, J.F., Gomila, A., Gaigg, S. B., Sivarajah, N. & Calvo-Merino, B. (2016). Dance Expertise modulates behavioural and psychophysiological responses to affective body movement. *Journal of Experimental Psychology*, 42(8), pp. 1139-1147. doi: 10.1037/xhp0000176

Fisher, V. J. (2017). Unfurling the wings of flight: clarifying 'the what'and 'the why'of mental imagery use in dance. *Research In Dance Education*, 18(3), 252-272.

Fox, R., & McDaniel, C. (1982). The perception of biological motion by human infants. *Science*, *218*(4571), 486–487. https://doi.org/10.1126/science.7123249

Gerson, S. A., Bekkering, H., & Hunnius, S. (2015). Short-term motor training, but not observational training, alters neurocognitive mechanisms of action processing in infancy. *Journal of Cognitive Neuroscience*, *27*(6), 1207–1214. https://doi.org/10.1162/jocn

Hespos, S. J., Grossman, S. R., & Saylor, M. M. (2010). Infants' ability to parse continuous actions: Further evidence. *Neural Networks*, *23*(8–9), 1026–1032. https://doi.org/10.1016/j.neunet.2010.07.010

Hespos, S., Saylor, M., & Grossman, S. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, *45*(2), 575–585. https://doi.org/10.1037/a0014145

Hofer, T., Hauf, P., & Aschersleben, G. (2005). Infant's perception of goal-directed actions performed by a mechanical device. *Infant Behavior and Development*, *28*(4), 466–480. https://doi.org/10.1016/j.infbeh.2005.04.002

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. https://doi.org/10.3758/BF03212378

Kilner, J. M., Marchant, J. L., & Frith, C. D. (2009). Relationship between activity in human primary motor cortex during action observation and the mirror neuron system. *PLoS ONE*, *4*(3), e4925. https://doi.org/10.1371/journal.pone.0004925

Kim, M., & Schachner, A. (2023). The origins of dance: Characterizing the development of infants' earliest dance behavior. *Developmental Psychology*, *59*(4), 691.

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, 39(1), 75-91.

Laban R von, Ullmann L. Modern educational dance. Macdonald and Evans; 1975.

Levine, D., Hirsh-Pasek, K., Pace, A., & Golinkoff, R. M. (2017). A goal bias in action: The boundaries adults perceive in events align with sites of actor intent. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(6), 916–927. https://doi.org/10.1037/xlm0000364

Monroy, C., Gerson, S. A., & Hunnius, S. (2018). Translating visual information into action predictions: Statistical learning in action and nonaction contexts. *Memory and Cognition*, 1–14. https://doi.org/10.3758/s13421-018-0788-6

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38. https://doi.org/10.1037/h0035584

Opacic, T., Stevens, C., & Tillmann, B. (2009). Unspoken knowledge: Implicit learning of structured human dance movement. *Journal of Experimental Psychology: Learning,* *Memory, and Cognition*, 35(6), 1570–1577. https://doi.org/10.1037/a0017244

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*(1), 155–184. https://doi.org/10.1016/j.cognition.2008.02.007

Papafragou, A., & Selimis, S. (2010). Event categorisation and language: A cross-linguistic study of motion. *Language and Cognitive Processes*, *25*(2), 224–260. https://doi.org/10.1080/01690960903017000

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, *129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Savaki, H. E., Kavroulakis, E., Papadaki, E., Maris, T. G., & Simos, P. G. (2022). Action observation responses are influenced by movement kinematics and target identity. Cerebral Cortex, 32(3), 490-503.

Schachner, A., & Carey, S. (2013). Reasoning about 'irrational'actions: When intentional movements cannot be explained, the movements themselves are seen as the goal. *Cognition*, *129*(2), 309-327.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition*. https://doi.org/10.1016/S0

Sharon, T., & Wynn, K. (1998). Individuation of actions from continuous motion. *Psychological Science*, *9*(5), 357–362. https://doi.org/10.1111/1467-9280.00068

Stahl, A. E., Romberg, A. R., Roseberry, S., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Infants segment continuous events using transitional probabilities. *Child Development*, *85*(5), 1821–1826. https://doi.org/10.1111/cdev.12247

Stapel, J. C., Hunnius, S., & Bekkering, H. (2012). Online prediction of others' actions: the contribution of the target object, action context and movement kinematics. Psychological research, 76, 434-445.

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257. https://doi.org/10.1037/a0015631

Talmy, L. (2000). Toward a Cognitive Semantics. In *Toward a Cognitive Semantics*. The MIT Press. https://doi.org/10.7551/mitpress/6848.001.0001

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of child language*, *19*(2), 311-333.

Wagner, L., & Carey, S. (2003). Individuation of objects and events: A developmental study. *Cognition*, *90*(2), 163–191. https://doi.org/10.1016/S0010-0277(03)00143-4

Woodward. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Wurm, M. F., & Schubotz, R. I. (2012). Squeezing lemons in the bathroom: Contextual information modulates action recognition. *NeuroImage*, *59*(2), 1551–1559. https://doi.org/10.1016/j.neuroimage.2011.08.038

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology*, *130*(1), 29–58. https://doi.org/10.1037/0096-3445.130.1.29

Zacks, Jeffrey, Kumar, S., Abrams, R., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216. https://doi.org/10.1016/j.cognition.2009.03.007

Zacks, JM. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008. https://doi.org/10.1016/j.cogsci.2004.06.003

Zacks, JM, Speer, N., Swallow, K., & Braver, T. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273. https://doi.org/http://dx.doi.org/10.1037/0033-2909.133.2.273

Figure Captions

Figure 1. Example frames from one dance video, with the frame depicting the end of each step outlined in red. Dance videos varied naturally in the timing of the four steps.

Figure 2. Response distribution across all participants, divided into 520ms bins. Colored bars indicate bins containing a 'true' step according to the dancers' intentions. *Denotes bins that were identified at high rates (> 1 *SD* above the mean frequency across all bins) ** = 2 *SD*.

Figure 3: Responses from Experiment 2 for a representative dancer. Each row depicts the frames selected for each response, with the corresponding label printed to the right. Colors indicate unique labels. Dotted lines indicate the step boundaries corresponding to the dancer's intentions. Graphs for dancers 2-6 can be viewed on https://osf.io/z4yu6/.

Figure 4: The relationship between how easy or difficult a step was to label in Experiment 2 and how successfully it was segmented relative to other steps in Experiment 1.