**Multiple meanings of complexity in the courtship displays of a lek-mating birds**

**ABSTRACT**

Studies of courtship displays, especially vocal displays in birds, have linked reproductive success to behavioral complexity. These studies often address a simple level of complexity—repertoire complexity—which focuses on e.g., the number of unique display elements in a display. More recent work has begun addressing another level—syntax complexity—which focuses on how organisms arrange the elements of their display. We analyze 353 displays of lek-mating Golden-winged Manakins, *Masius chrysopterus,* by coding their dances as strings of behavioral elements. In addition to a basic measure of repertoire complexity (number of unique elements in a display), we use methods from information science to calculate multiple measures of syntax complexity (entropy and compressibility). We find that solo male displays (SOLO) were the most complex, whereas unsuccessful displays performed for a female audience (AUDI) were of intermediate complexity, and successful displays ending in copulation (COP, n = 13) were the simplest in terms of both repertoire and syntax. Using Jaro string distance, a method from record-linkage theory, we find that COP displays had the most uniform syntax and that displays varied more between context (SOLO, AUDI, COP) than between individual. Females chose to copulate after simple displays building to a dramatic high-speed element (the audible log-approach dive). These results raise questions about the difference between song and dance displays, the role of audience participation, and a third level of complexity—element complexity—that reckons with the aesthetic content of behaviors themselves.

**KEYWORDS**

complexity, repertoire, syntax, courtship, display element, lossless compression, entropy, Jaro string distance, lek, manakin

## **INTRODUCTION**

One of the oldest puzzles in evolutionary biology is the process by which elaborate, sexual display traits and behaviors evolve ($Darwin\_Descent, $Fisher, $Kirkpatrick\_1982, $Prum). Recent work has sought to explain the diversity and elaboration of sexual displays by conceiving of complexity itself as a character (Byers & Kroodsma, 2009; Bradbury & Vehrencamp, 2014; Patricelli & Hebets, 2016). One perspective is that complexity acts as an honest signal of viability, whereby the strongest, smartest, or healthiest potential mates also perform the most complex displays ($CITE)

Behavioral complexity as a locus of mate choice is particularly well-studied in bird songs. Early work focused on repertoire complexity, or the simple number of behavioral elements contained in a display. For example, female $EX and $EX prefer males who sing songs with large repertoires. Recent work has made an important step beyond repertoire complexity to syntax complexity, or the different ways a given repertoire can be arranged. Syntax complexity is often measured in terms of entropy (sensus $SHANNON). A low-entropy process produces predictable patterns, whereas a high-entropy process produces more “surprising,” less predictable, arrangements. For example, female $EX and $EX prefer males who sing higher entropy sequences of songs.

Entropy characterizes a probabilistic process producing a display, but related metrics can help characterize and compare displays themselves. The methods of compression, readily drawn from computer science, are used to condense data based on common motifs or repeated stretches of elements ($CITE). In other words, a display with a simple or repeated syntax will be highly compressible. Another method from record-linkage theory, Jaro distance, can be used to directly compare the syntax of different displays. Jaro (1989) developed a simple but elegant method for matching census records from disparate sources. The algorithm analyses the number of transpositions and mismatches between two strings of data and assigns a distance between 0 (no matches) and 1 (complete match). Where high compressibility indicates a display with a simple syntax, low Jaro distance indicates two displays with a similar syntax.

Here, we investigate the complexity of courtship displays in Golden-winged Manakins (*Masius chrysopterus*), a tropical lekking bird. Manakins (Aves, Pipridae) are birds well-known for their extraordinary courtship displays and ornate plumage (Kirwan and Green $CITE). With the highest percentage of lek-mating species (~84%), manakins are also among the “lekkiest” families in the animal kingdom (McDonald 2010). Lek mating systems are particularly useful for examining sexual displays because males provide neither parental care nor other resources valuable to breeding females (Gibson & Bradbury, 1985). Females, therefore, choose mates based solely on ornamentation (Zuk et al., 1990), the characteristics of physical courtship displays (Barske, Schlinger, Wikelski, & Fusani, 2011), or a combination of elaborate plumage and display (Kodric-Brown & Nicoletto, 2001).

Golden-winged Manakinsperform elaborate, highly acrobatic displays that remain little-studied (Prum & Johnson, 1987, Snow & Snow 1992). We used video observation to capture and characterize hundreds of courtship displays across multiple males and display sites. Our dataset included displays in three distinct contexts: solo male displays (SOLO); unsuccessful displays for an audience (AUDI); and successful displays for an audience that ended in copulation (COP). We asked whether COP displays were more complicated than either AUDI or SOLO displays in terms of repertoire complexity (number of unique behavioral elements), syntax complexity (entropy, compressibility), and syntax comparisons (Jaro distances across contexts and individuals). We bring our results to bear on broader questions about sexual displays and preferences.

## **METHODS**

### *Study Site*

We collected data in 2015-2016 (January, June-August) and 2017 (September-December) at the Milpe Bird Sanctuary of the Mindo Cloudforest Foundation, in northwestern Ecuador (~0°1’48”N, 78°57’12” W). Milpe comprises 100 ha of west slope Chocó-Andean forest (1,100 m elevation) with a network of maintained trails.

*Study Species*

Because the Golden-winged Manakinis currently monotypic in their genus, we refer to it by the generic name *Masius*. Male *Masius* tend small sections of fallen mossy logs (~20–60 cm2 surface) on which they perform elaborate gymnastic displays ($PRUM). Display logs are often in aural, but rarely in visual, contact with one another as part of a broader, dispersed lek (Bradbury, 1981). As in many other manakin species, female *Masius* have an overall green plumage while males undergo a process of plumage maturation from predefinitive (green) to definitive (black, golden-yellow, and/or orange-red) plumages over multiple years (Taylor et al. 2020, Schaedler et al. 2021). Some older predefinitive-plumaged males can be identified by waxy nape feathers, a golden horn-like crest, or golden forehead feathers. Females are only known to copulate with definitive-plumage males (*pers. obs.*).

### *Field Methods*

We used mist-nets (6 or 12 m length, 30mm mesh) to capture and mark individual birds. All *Masius* were banded with a numbered aluminum band and a unique combination of plastic color bands. Other species were released immediately. All field methods were approved by the University of Wyoming Institutional Animal Use and Care Committee (Protocol#20160602DM00242-02).

After regular observations to identify *Masius* display logs, we monitored logs with video cameras. We prioritized video surveillance at logs with high manakin activity or female visitation, or at snewly discovered logs. Individual cameras (Sony Handycam HDR-CX405 or Sony Handycam HDR-CX240, Sony Corp., Tokyo, Japan) were housed in weatherproof container and powered with an external batter (12v motorcycle battery or 10,000 mAh GETIHU power bank, Shenzhen Top Star Industry Co. Ltd., China). Recording ended when the memory card reached capacity, the batteries died, or weather necessitated camera retrieval. We ran 4 to 6 cameras daily, one per log, with each camera recording approximately 7 hrs of video. Footage was filtered with a liberal motion detection program ($GITHUB) using the OpenCV library ($CITE) in Python v. ($CITE). After being flagged by the motion detection program, motion clips were manually verified and bookmarked for subsequent coding.

*Displays and behavioral elements*

We defined a display as a sequence of distinct behavioral elements during which time males were never absent for >60 s. We only included sequences that lasted >60 s and included two core behavioral elements, “Audible log-approach dive” and “Side-to-side bow” (Table 1; $PRUM\_JOHNSON, $TAYLOR).

*Masius* displays can involve multiple males and females both as performers and audience members ($PRUM\_JOHNSON, Taylor et al. 2018). For this study, we excluded all displays featuring multiple dancing males (n = 26) or a single, identified male audience (n = 3, all predefinitive Male #980, including one copulation). We retained 21 displays where an audience member was suspected as a predefinitive male based on plumage but performed no display behavior. All displaying males in our final dataset had definitive plumage.

We categorized the remaining 353 male displays into three contexts: SOLO, AUDI, and COP. SOLO displays were solo male performances, AUDI displays featured one or two audience members at some point during the display but did not end in successful copulation, and COP displays ended in a successful copulation. For our main analyses, we ended COP displays at first copulation. A comparison of before- and after-copulation displays is given in Supplementary Material.

We coded display elements using Behavioral Observation Research Interactive Software (BORIS), an open-source event logging tool (Friard & Gamba, 2016). A total of 41 elements occurred in the raw BORIS logs (Table S1). We subsequently excluded tracking elements (e.g., “Start”), movement-based elements (e.g., “Male1 On Log”, “Female Movement”), female responses (e.g., “Female Tracking Male”), and male behaviors not directly involved in display dances (e.g. “Vocalization,” “Gardening”). Finally, we retained only one half of paired elements (e.g., excluding “Side-to-side bow Left” while keeping “Side-to-side bow Right” as “Side-to-side Bow”) and excluded “Attempted copulation” and “Copulation” elements. The 12 remaining, core display behaviors—including pauses (“Zero”)—are described in Table 1.

### *Repertoire complexity*

We calculated three metrics related to the repertoire of individual *Masius* displays: (A) Duration, in seconds; (B) Length, in number of total elements; and (C) Repertoire size, in number of unique elements. Display duration was calculated from the raw data and thus included the timing of some elements excluded from other behavioral analyses.

We compared these metrics across display context (SOLO, AUDI, COP) with ANOVA and Tukey’s HSD. To address our small sample size of COP displays (n = 13), we compared our COP metrics to a randomized distribution. Across each of 10,000 replicates, we randomly selected (without replacement) 13 displays from any context in our full dataset. We then compared the distribution of randomly-selected display values to empirical COP values. All analyses were conducted with the *tidyverse* packages in Program R v4.2.2 ($TIDYVERSE, R).

*Syntax complexity*

We quantified syntax complexity of displays with two related measures: entropy and compressibility. We began by translating displays into strings of individual characters (Table 1). We computed the first-order entropy display strings with *entropy* function in R package *acss* ($CITE). The resulting values depend on the empirical frequency of each element in each individual display. We thus scaled each value by the maximum possible entropy for a display—given as log2(number of unique elements)—giving a final metric between 0 and 1 (Vanderbilt, Kelley, & DuVal 2015). We computed compressibility using a standard text compression algorithm (LZ77 and Huffman Coding) implemented with default parameters in the R package *brotli* ($CITE). The final compressibility metric was taken as the ratio of the length of the uncompressed string to the compressed string.

We compared syntax complexity measures across context (SOLO, AUDI, COP) with ANOVA and Tukey’s HSD. As with repertoire complexity, we used a randomization procedure to compare the small sample of 13 COP displays with a randomized distribution of both entropy and compressibility values (10,000 random sets of 13 displays from the full dataset, drawn with no replacement).

Entropy and compressibility are fundamentally intertwined metrics. In theory, high-entropy systems produce outputs that lack well-defined, repeated motifs and are therefore difficult to compress ($CITE). We assessed the empirical correlation between entropy and compressibility in our data and investigated the variation diverging from that correlation.

*Context vs. individual variation*

We investigated how display syntax varied in terms of context (SOLO, AUDI, COP) or in terms of male individuality using Jaro string distances. We calculated the Jaro distance between all pairs of display strings with the R package *stringdist* ($CITE). Using only displays from identified males in each context, we then calculated four sets of distances from the focal display: (A) same-male/same-context, (B) same-male/different-context, (C) different-male/same-context, and (D) different-male/different context.

We used a randomization procedure to investigate two key, COP-related comparisons. First, we compared different-male/same-context Jaro distances from COP displays (n = 39 distances) to same-male/different-context distances from COP displays (n = 1,740). Second, we compared different-male/same-context distances from COP displays (n = 39) to different-male/same-context distances from AUDI (n = 2,444) plus SOLO (n = 19,475) displays. For both, we drew 10,000 random sets of 39 distances, without replacement, from the comparison dataset.

## **RESULTS**

*Displays and behavioral elements*

Our final dataset includes 353 *Masius* displays (252 across 15 display logs (1-102 displays per log). Of these displays, 341 were performed by one of 10 identified males (2-172 displays each, Table S2). All displays occurred between 24-Jun and 14-Jan, with the bulk of displays (278/353) between October and December (Fig. S1). The earliest AUDI display was 30-Jun and the earliest COP display was 26-Oct. The dataset featured 36 AUDI displays with one of 11 identified females (1-7 attendances each) and 5 COP displays with one of 3 identified females (1-2 copulations each; Table S3). All 13 COP displays were performed by one of 3 identified males (1-9 copulation each; Table S4).

By definition, every display in our dataset featured at least one Audible log-approach dive and Side-to-side bow. Representation of the remaining behaviors differed in terms of display context (Table 1, see Table S5 for raw frequencies). Silent log-approach dives were frequent across SOLO displays but present in only 1 AUDI display and 0 COP displays. To-and-fro flights were common in AUDI displays (77/89 displays) but relatively rare in SOLO (63/251) and COP (2/13). Neck twists were common in AUDI (86/89) and COP (9/13), but rare in SOLO (21/251). Head-down bows were nearly universal in SOLO (248/251) and AUDI (88/89) displays, but absent from COP. However, there were Head-down bows and the rare Metronome element in 2 and 1 after-copulation displays, respectively (Supplementary Material). Pauses, Mixed, and Other behaviors were more common across SOLO displays than AUDI and COP displays (Table 1).

*Repertoire complexity*

In terms of raw display duration, AUDI displays (mean ± S.D.: 217 ± 130 s) were significantly longer than both SOLO displays (136 ± 65 s) and COP displays (126 ± 41 s; overall ANOVA *P* < 0.001; Fig. 1A). In terms of the total number of display elements, AUDI displays (101 ± 72 elements) were again significantly longer and more variable than SOLO display (62 ± 16 elements), which in turn were significantly longer than COP displays (24 ± 13 elements; overall ANOVA *P* < 0.001; Fig. 1B).

In contrast to these patterns in display length, COP displays had significantly smaller repertoires (3.2 ± 0.8 unique elements) than the similar SOLO displays (5.9 ± 1.1 unique elements) and AUDI displays (5.9 ± 1.0 unique elements; overall ANOVA *P* < 0.001; Fig. 1C). Across 10,000 replicates of 13 randomly-drawn displays from our dataset, mean repertoire size was never less than the empirical mean of the 13 COP displays (Fig. S2).

*Syntax complexity*

SOLO displays showed significantly higher scaled entropy values (0.87 ± 0.07) than AUDI displays (0.71 ± 0.12), which in turn had significantly higher scaled entropy than COP displays (0.26 ± 0.16; overall ANOVA *P* < 0.001; Fig. 2A). There was an identical pattern in compressibility, measured as the ratio of uncompressed to compressed display string length. SOLO display strings had significantly lower compression ratios (1.07 ± 0.30) than AUDI display strings (2.80 ± 1.15), which were significantly less compressible than COP display strings (3.65 ± 1.01; overall ANOVA *P* < 0.001; Fig. 2B). None of 10,000 random sets of 13 displays for each metric had a lower mean entropy or higher mean compression ratio than the empirical set of 13 COP displays (Fig. S2).

As expected, entropy and compressibility were significantly correlated across our dataset (linear regression, adjusted *R2 =* 0.53, *P* < 0.001;Fig. 2C). However, wide variation highlighted the differences between these metrics. For example, the most compressible display string (AUDI ID-1487, compression ratio = 7.45) had intermediate entropy (0.63) but was long, with 365 total elements, and primarily made up of long stretches of Side-to-side bows and Neck twists. A display with similar entropy but much lower compression ratio (AUDI ID-453, scaled entropy = 0.64, compression ratio = 2.53) was shorter, with 81 elements, and featured a tail of individual behaviors (coded “*IEDBA*”; Table 1) that made it difficult to compress. Indeed, our choice of compression algorithm (LZ77 and Huffman Coding) and our final compression ratio metric created an overall positive correlation between display length and compressibility (Fig. S3).

The lowest entropy display (COP ID-1533, scaled entropy = 0.11, compression ratio = 4.71) was simply 66 Side-to-side bows followed by an Audible log-approach dive. A closer look revealed a characteristic syntax for COP displays (Table S2). Ten of 13 COP displays, across all 3 copulating males, ended with long (34+) stretches of Side-to-side bows followed by an Audible log-approach dive. One display (ID-1455) was similar except for the insertion of a single Half-bow before the dive, while another (ID-1987) had a short sequence of two additional elements—To-and-fro flights and then a Neck twist—before the final bow and dive combination. The remaining display (ID-1824) was primarily stretches of Neck twists, yet still ended with a bow and dive combination (Table S2).

*Context vs. individual variation*

Judged by average Jaro distances, displays in every context (SOLO, AUDI, COP) were more similar to displays of the same context than to displays of other contexts, regardless of the individual male performing the display (Fig. 3). Within each context, the distances among displays by the same male (mean ± SD, 0.37 ± 0.11) was only slightly lower than among displays by different males (0.39 ± 0.11). In contrast, displays in different contexts were more distant whether given by the same male (0.53 ± 0.10) or different males (0.54 ± 0.10). The mean distance for COP *vs.* different-male/same-context displays (n = 39 distances) was lower than the mean of all 10,000 random draws (each n = 39) of COP *vs.* same-male/different context distances (Figure S4).

Within each context, mean Jaro distance among displays from different males was shorter for COP displays (0.14 ± 0.10) than for either AUDI displays (0.36 ± 0.10) or SOLO displays (0.39 ± 0.11). The mean distance for COP *vs.* different-male/same-context distances (n = 39) was lower than the mean of all 10,000 random draws (each n = 39) of AUDI or SOLO *vs.* different-male/same-context displays (Figure S4).

Average similarity comparisons were not absolute rules. Across the 13 COP displays, 6 were closest (i.e., lowest Jaro distance) to another COP display by the same male and 3 to another COP display by a different male. However, 2 COP displays were closest an AUDI display by the same male and 2 to an AUDI display by a different male.

## **LITERATURE CITED**