**Female audiences compose male courtship displays in a lek-mating bird**

**ABSTRACT**

Animal behavior research usually assumes the characteristics of courtship displays, such as quality or complexity, are intrinsic traits of the organism performing the display. Here we provide evidence that the composition and complexity of courtship displays are determined by the female audience, rather than the male performer, in lek-mating Golden-winged Manakins (*Masius chrysopterus*). We use video observation to analyze 423 male courtship displays in three audience contexts: Solo displays with no audience (SOLO, n = 308), unsuccessful displays for females (AUDI, n = 108), and rare, successful displays for females ending in copulation (COP, n = 13). Using entropy and a novel metric we call compressibility, we find the arrangement of behavioral elements is least predictable and repetitive in SOLO displays, more so in AUDI displays, and most in COP displays. Using Jaro string distance, a method from record-linkage theory, we directly quantify similarities in display arrangement. Displays vary more by audience context than individual male performer, and COP displays are arranged most uniformly. Finally, we show males perform different behaviors based on the location and behavior of their female audience. We find no evidence that males are intrinsically constrained or consistent in arranging their displays. Our results suggest the courtship displays of Golden-winged Manakins are joint compositions by females and males, not merely controlled demonstrations of male traits. Future studies of sexual selection and mate choice must distinguish features of organisms (e.g., genetic traits or preferences) from features of interactions between organisms.

**KEYWORDS**

complexity, entropy, Jaro string distance, lek, lossless compression, manakin, mate choice, repertoire, syntax

## **INTRODUCTION**

Studies of sexual selection and animal behavior traditionally ask how performer traits (e.g., a peacock’s tail) influence audience behavior (e.g., a peafowl’s decision to copulate; Darwin, 1871; Prum, 2017). Quantitative models of sexual selection demonstrate how audience mating preferences can result in the evolution of elaborate, intrinsic performer traits, even when those traits are costly to develop and maintain (e.g., Fisher, 1930; Kirkpatrick, 1987; Kokko et al., 2003; Lande, 1981). Inspired by theory, a vast amount of empirical work in animals investigates how audiences judge intrinsic performer traits during mate choice (Andersson, 1994; Rosenthal, 2017).

However, some studies of animal courtship suggest we cannot understand mate choice strictly in terms of audience preferences for intrinsic performer traits. For example, male Satin Bowerbirds (*Ptilonorynchus violaceus*) decrease the intensity of their movement displays in response to female crouching signals (Patricelli et al., 2002), and male fiddler crabs (*Austruca perplexa*) change the frequency of claw-waving displays when females approach (How et al., 2008). In these species, a female cannot objectively judge movement intensity or claw-waving frequency as intrinsic male traits; those characteristics change with respect to the expression of her own preferences.

We ask whether female audiences or male performers are more important in determining the repertoire, complexity, and composition of courtship displays in a tropical, lek-mating bird, the Golden-winged Manakin (Passeriformes: Pipridae: *Masius chrysopterus*, monotypic in the genus and hereafter *Masius*). Male *Masius* perform acrobatic courtship displays to entice prospective female mates, thereafter providing no parental care (Prum & Johnson, 1987; Snow & Snow, 1992). We used video observations to characterize hundreds of *Masius* displays in three audience contexts: solo male displays (SOLO), unsuccessful displays for a female audience (AUDI), and successful displays for a female audience that ended in copulation (COP).

First, we show how display characteristics—duration, repertoire size, entropy, and compressibility (a measure of simplicity and repetition)—vary across audience contexts, with COP displays being the simplest and most repetitive. Second, we use a novel application of record-linkage theory (Jaro string distance; Fellegi & Sunter, 1969; Jaro, 1989) to show displays are more similar when performed by different males in the same audience context than when performed by the same male in a different audience context. Third, we show how simple female behaviors directly influence the behavioral elements in male displays. These results indicate *Masius* courtship displays are dynamic interactions between audiences and performers, rather than straightforward demonstrations of intrinsic performer traits under sexual selection.

## **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*

Male *Masius* tend small sections of fallen, mossy, slanted logs (~20–60 cm2 surface) on which they perform gymnastic displays (Prum & Johnson, 1987). 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 (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, 30 mm 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 newly 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 flagged with a liberal motion detection program (<https://github.com/ltaylor2/Masius_Movement>) using OpenCV v3.3 (Bradski, 2000) in Python v2.7. Motion clips were then 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 included only sequences that lasted >60 s and included at least one of two core behavioral elements: “Audible log-approach dive” or “Side-to-side bow” (Prum & Johnson, 1987; Taylor et al., 2020).

*Masius* displays can involve multiple males and females both as performers and audience members (Prum & Johnson, 1987; Taylor et al., 2020) For this study, we excluded all displays featuring multiple dancing males (n = 26) and displays with only male audiences (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 423 male displays into three contexts: SOLO, AUDI, and COP. SOLO displays were solo male performances, AUDI displays featured a female audience member 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 behavioral elements—including pauses (“Zero”)—are described in Table 1.

### *Display Characteristics*

We calculated five metrics related to organization and complexity of *Masius* courtship displays: (A) Duration, in seconds; (B) Length, in number of total elements; (C) Repertoire size, in number of distinct elements; (D) Entropy; and (E) Compressibility. Duration was calculated from the raw data and thus included the timing of some elements excluded from other behavioral analyses. Whereas repertoire size measures a display in terms of the number of behavioral elements (Catchpole, 1987; Nowicki & Searcy, 2004), entropy and compressibility measure syntax, i.e. the arrangement of a given repertoire of elements. First-order entropy (Shannon, 1948) is a common measure of syntax complexity in animal behavior (e.g., Da silva et al., 2000; Freeberg & Lucas, 2012; McCowan et al., 2002; Palmero et al., 2014; Vanderbilt et al., 2015). A low-entropy system is associated with predictable patterns, whereas a high-entropy system is associated with greater “surprise.” However, entropy is more technically a probabilistic property of a source (e.g., a system that produces a display) than a direct property of a message (e.g., a display itself). To characterize display syntax more directly, we thus also used compression methods from computer science, which condense data based on redundant elements or repeated motifs (Huffman, 1952; Ziv & Lempel, 1977). A display with a simple or repeated syntax will be highly compressible. High compressibility is only possible in low entropy systems (Shannon, 1948), meaning these two metrics are not independent of one another.

To calculate entropy and compressibility, we translated displays into strings, where each kind of behavioral element was a unique character (Table 1). We computed first-order entropy from each display string using the *entropy* function in R package *acss* (Gauvrit et al., 2016). The resulting entropy 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 distinct elements)—for a final metric between 0 and 1 (Vanderbilt et al., 2015). We computed compressibility using a standard text compression algorithm (LZ77 and Huffman Coding) implemented with default parameters in the R package *brotli* (Ooms & Google, Inc., 2022) . Our final compressibility metric was compression ratio, or the length of the uncompressed string divided by the compressed string.

We compared our five metrics across display contexts (SOLO, AUDI, COP) with linear models, using AUDI as context intercept and including male ID and observation month as discrete fixed effects. Unidentified males were assumed to be unique in each display. To address our small sample size of COP displays (n = 13), we compared COP metrics to a randomized distribution. Across each of 100,000 replicates, we randomly selected (without replacement) 13 displays from any context in our full dataset. We then compared the distribution of randomly-selected metrics to empirical COP metrics. All analyses were conducted with the *tidyverse* packages in Program R v4.2 (R Core Team, 2022; Wickham et al., 2019).

*Display Similarity: Male performer vs. Audience context*

We investigated whether the syntax of displays varied more in terms of audience context (SOLO, AUDI, COP) or in terms of male individuality using Jaro string distances. Jaro (1989) developed an 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). Low Jaro distance indicates two displays have a similar syntax.

We calculated the Jaro distance between all pairs of display strings with the R package *stringdist* (van der Loo, 2014)*.* Using only displays from identified males, we calculated four sets of distances between displays: (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 confirm two 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 = 2,030). Second, we compared different-male/same-context distances from COP displays (n = 39) to different-male/same-context distances from AUDI (n = 3,289) plus SOLO (n = 29,648) displays. For both comparisons, we drew 100,000 random sets of 39 distances, without replacement, from the comparison dataset.

*Female Behavior*

We asked how simple female behaviors directly influenced male behavioral elements during the composition of a courtship display. For each male behavioral element in the dataset, females were marked as on or off the display log. When on the display log, females were marked as either above or below the male (i.e., upslope or downslope). We quantified the behavioral elements performed by males when females were on/off the log and above/below the displaying male for both AUDI and COP displays. Elements were excluded when female location was not coded.

## **RESULTS**

*Displays and Behavioral Elements*

Our final dataset includes 423 *Masius* displays across 15 display logs (1-109 displays per log), with 308 SOLO, 102 AUDI, and 13 COP displays. Across all displays, 406 were performed by one of 10 identified males (1-199 displays each, Table S2). All displays occurred between 24-Jun and 14-Jan, with 77% of displays (324/423) occurring between October and December (Fig. S1). The earliest AUDI display was 30-Jun and the earliest COP display was 26-Oct. The dataset featured 33 AUDI displays with one of 10 identified females (1-8 attendances each) and 4 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 or 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 (84/102 displays) but relatively rare in SOLO (85/308) and COP (2/13). Neck twists were common in AUDI (97/102) and COP (9/13), but rare in SOLO (27/308). Head-down bows were in the vast majority of SOLO (289/308) and AUDI (89/102) 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 behaviors, and Other behaviors were more common across SOLO displays than AUDI and COP displays (Table 1).

*Display Characteristics*

All five display characteristics—duration, length, repertoire size, entropy, and compressibility—varied significantly across audience contexts (Fig. 1). In terms of display duration, AUDI displays (mean ± SD: 204 ± 128 s) were significantly longer than both SOLO displays (133 ± 63 s) and COP displays (125 ± 41 s; model estimate AUDI *vs*. SOLO: -82, *P* < 0.001; AUDI *vs.* COP: -79, *P* < 0.001; overall *P* < 0.001, adjusted *R2­* = 0.18; Fig. 1A). In terms of display length, AUDI displays (94 ± 70 elements) were significantly longer than both COP displays (62 ± 16 elements) and SOLO displays (23 ± 13 elements; AUDI *vs.* SOLO: -76, *P* < 0.001; AUDI *vs.* COP: -37, *P* < 0.001; overall *P* < 0.001, *R2­* = 0.42; Fig. 1B). COP displays had significantly smaller repertoires (3.2 ± 0.8 distinct elements) than AUDI displays (5.5 ± 1.3 distinct elements), the latter of which did not differ from SOLO displays (5.6 ± 1.2 distinct elements; AUDI *vs.* SOLO: 0.1, *P* = 0.32; AUDI *vs.* COP: -2.3, *P* < 0.001; overall *P* < 0.001, *R2­* = 0.15; Fig. 1C). Across 100,000 replicates of 13 randomly-drawn displays, mean repertoire size was never less than the empirical mean of the 13 COP displays (Fig. S2).

SOLO displays showed significantly higher entropy (0.87 ± 0.08) than AUDI displays (0.70 ± 0.16), which in turn had significantly higher entropy than COP displays (0.26 ± 0.16; AUDI *vs.* SOLO: 0.17, *P* < 0.001; AUDI *vs.* COP: -0.44, overall *P* < 0.001, *R2­* = 0.58; Fig. 1D). 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 (1.04 ± 0.31) than AUDI display strings (2.73 ± 1.14), which were significantly less compressible than COP display strings (3.65 ± 1.01; AUDI *vs.* SOLO: -1.77, *P* < 0.001; AUDI *vs.* COP: 0.87, *P* < 0.001; overall *P* < 0.001, *R2­* = 0.63; Fig. 1C). None of 100,000 random sets of 13 displays had lower mean entropy or higher mean compression ratio than the empirical set of 13 COP displays (Fig. S2).

There were additional significant effects of observation months and individual males in for some metrics (Supplementary Material). As dictated by theory (Shannon, 1948) entropy and compressibility were significantly correlated across our dataset (linear regression, *P* < 0.001, adjusted *R2 =* 0.46;Fig. S$). 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).

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+ element) 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).

*Display Similarity: Male performer* vs. *Audience context*

Using Jaro distance, displays in every context (SOLO, AUDI, COP) were on average more similar to displays of the same context than to displays of other contexts, regardless of the male performing the display (Fig. 2). Although increasing Jaro distance was positively correlated with differences in display length or the number of distinct elements, these correlations explained only a small fraction of variance (Fig. S4).

Displays performed by the same male in the same context were on average slightly more similar to one another (mean Jaro distance ± SD: 0.40 ± 0.13) than displays by different males in the same context (0.41 ± 0.13). In contrast, displays in different contexts were on average less similar to one another whether performed by the same male (0.55 ± 0.13) or different males (0.57 ± 0.14). The mean distance for COP *vs.* different-male/same-context displays (n = 39 distances) was lower than the mean of all 100,000 random draws (each n = 39) of COP *vs.* same-male/different context distances (Fig. S5).

On average, COP displays performed different males were more similar to one another (mean Jaro distance ± SD: 0.14 ± 0.10) than either AUDI displays (0.36 ± 0.11) or SOLO displays (0.41 ± 0.13). The mean distance for COP *vs.* different-male/same-context distances (n = 39) was lower than the mean of all 100,000 random draws (each n = 39) of AUDI or SOLO *vs.* different-male/same-context displays (Fig. S5).

Average similarity comparisons were not absolute rules. Of the 13 COP displays, 5 had the lowest Jaro distance to a COP display by the same male, and 2 to a COP display by a different male. However, 4 COP displays were most similar to an AUDI display by the same male and 2 to an AUDI display by a different male.

*Female behavior*

**DISCUSSION**

Our major finding is that Masius courtship displays become steadily simpler as they progress from those performed with no female audience (SOLO), to those performed for one or more females but not resulting in a copulation (AUDI), to those resulting in a copulation (COP). This progressive simplicity is best captured by measures of syntactic complexity (entropy and compression ratio, which emphasize not just element quantity, but their arrangement (Fig. 2). In contrast, repertoire complexity (Fig. 1) more typical of traditional analyses of avian vocal repertoires captures only the low number of distinct display elements characterizing COP displays. Further, not only are COP displays syntactically simpler (a few elements arranged in a simple, orderly arrangement) but the displays of different males are more similar to each other than the displays of individual males are to the rest of their own COP displays. That across-male similarity suggests that female preference drives male performance to a standard, simple and orderly arrangement of a few distinct display elements – a long series of Bows followed by Neck twists and audible log-approach dives – regardless of male or female identity.

A fundamental feature of the natural sciences, from physics to chemistry to biology, is that the arrangement of elements is as important as the ingredients. For example, water and ice have the same ingredients but differ dramatically in their properties. We have shown that sequences of courtship display elements have a similar strong dependence on the arrangement of elements – the order matters. We show that as context varies, from displays performed by solo males (SOLO), to displays performed for a female audience (AUDI), to successful displays followed by a copulation (COP), the arrangement of display elements (what we call syntax complexity) becomes simpler, as assessed by entropy and compressibility (Fig. 2). Further, using Jaro string distance, a method from record-linkage theory, we find that successful displays preceding a copulation had the most uniform syntax and that displays varied more between the social contexts (female presence and prior to a copulation) than between individuals. We conclude that using novel tools for assessing the arrangement of behavioral elements (syntax complexity and Jaro distance) can help elucidate the factors that drive the evolution of elaborate behaviors involved in sexual selection.

## **LITERATURE CITED**