Reliability of meta-analyses in ecology and evolution: (mostly) good news from a case study on sexual signals

Pietro Pollo[1]*, Malgorzata Lagisz[1], Renato Chaves Macedo-Rego[2], Ayumi Mizuno[1,3], Yefeng Yang[1], Shinichi Nakagawa[1,3]

[1] Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, Kensington, NSW, 2052, Australia

[2] Departamento de Biologia Geral, Universidade Federal de Viçosa, Av. Peter Henry Rolfs, Viçosa, Minas Gerais, 36570-000, Brazil

[3] Department of Biological Sciences, University of Alberta, CW 405, Biological Sciences Building, Edmonton, AB T6G 2E9, Canada

* Corresponding author: pietro_pollo@hotmail.com

Abstract

Meta-analyses are powerful synthesis tools that are popular in ecology and evolution due to the rapidly growing literature of this field. Although the usefulness of meta-analyses depends on their reliability, such as the precision of individual and mean effect sizes, attempts to reproduce meta-analyses' results remain rare in ecology and evolution. Here, we assess the reliability of 41 meta-analyses on sexual signals by evaluating the reproducibility and replicability of their results. We attempted to (1) reproduce meta-analyses' mean effect sizes using the datasets they provided, (2) reproduce meta-analyses' effect sizes by re-extracting 5,703 effect sizes from 246 empirical studies they used as sources, (3) assess the extent of relevant data missed by original meta-analyses, and (4) replicate meta-analyses' mean effect sizes after incorporating re-extracted and relevant missing data. We found many discrepancies between meta-analyses' reported results and those generated by our analyses for all reproducibility and replicability attempts. Nonetheless, we argue that the meta-analyses we evaluated are largely reproducible and replicable because the differences we found were small in magnitude, leaving the original interpretation of these meta-analyses' results unchanged. Still, we highlight issues we observed in these meta-analyses that affected their reliability, providing recommendations to ameliorate them.

Keywords: data integrity, methodological accuracy, open science, research synthesis, sexual selection.

Introduction

The literature of ecology and evolution, like of other fields of study, is expanding rapidly [1,2]. Consequently, synthesising this growing body of work becomes increasingly necessary to identify patterns across individual studies. Meta-analyses, which are perceived as the gold standard for evidence synthesis, can contribute to this endeavour because they aim to detect and retrieve all relevant studies on a given topic, extract data from these studies, and quantify an average effect of interest using the extracted data [3,4]. Nonetheless, many challenges arise during the execution of meta-analytical studies, requiring researchers to be vigilant to ensure their reliability.

Foremost, meta-analyses need to be transparent, which means that they must provide details on decisions and resources regarding all steps of their implementation (e.g. searches, screening, data extraction, analysis code, etc). Yet, recent appraisals of the secondary literature in ecology and evolution show that meta-analyses in this field are often poorly transparent [5–7]. Reporting guidelines, such as MOOSE [8] and PRISMA [6,9], were developed to address transparency issues in evidence syntheses. Researchers who adhere to these guidelines can thus enhance the quality of their meta-analyses. Still, transparency represents only the first step for high-quality syntheses as providing information does not guarantee its reliability. For instance, researchers can make mistakes when describing their methods, provide faulty code for their analyses, or even forge data, impairing the reproducibility and replicability of their findings. Therefore, it is imperative to evaluate studies beyond transparency.

Meta-analyses possess the advantage of having sources of information that are readily accessible (i.e. individual studies from which data are extracted). This means that meta-analyses can be more easily evaluated than other approaches regarding reproducibility and replicability aspects. More specifically, meta-analytical studies that report their decision criteria (i.e. studies and data considered valid for inclusion) and their effect sizes with information on their sources (i.e. studies they were extracted from) should allow evaluations regarding the reproducibility of their dataset and their results. However, to our knowledge, only computational reproducibility of meta-analyses (i.e. reproducibility of their code) has been examined in ecology and evolution [10]. By contrast, reproducibility attempts of data extraction and general results have only been conducted in other fields of study (but see [11]; e.g. medicine [12]; psychology [13]). Furthermore, other elements connected to the reliability of meta-analyses, such as their efficacy in detecting relevant studies, remain virtually unexplored.

Here, using 41 meta-analyses related to sexual signals [14–54], we conduct the largest reproducibility and replicability effort for meta-analyses ever done in terms of number of re-extracted primary studies and effect sizes (5,703 data points from 246 empirical studies, representing almost a fifth of all primary studies in our dataset). We evaluate multiple reliability aspects related to distinct implementation stages of these meta-analyses (Figure S1). First, we re-analyse the dataset provided by these meta-analyses, comparing the mean effect sizes we obtained with those reported in them. Second, we extract data from their original sources (i.e. individual studies) and compare these extracted data points with the ones reported in these meta-analyses. Third, we assess how many data points from the original sources that we verified should have been extracted and included in these meta-analyses' datasets but were not (i.e. "missed" data). Fourth, because these meta-analyses ask similar questions involving sexual signals, we estimate the minimum number of studies that contained relevant data but were not listed as sources in these meta-analyses (i.e. undetected studies). Fifth, we assess the extent that the results of these meta-analyses change when re-extracted data (along with missed data and undetected studies) is analysed in place of originally reported data.

Methods

This manuscript is part of a larger research project that uses data from specific meta-analyses (see subsections below). Our methodology, summarised in Figure S1, was described in our pre-registration [55], and we adhered to it as much as possible (see changes in Supplementary information S1). We report author contributions using MeRIT guidelines [56] and the CRediT statement [57].

Reported dataset

A recent systematic map identified the existence of 151 meta-analyses on topics related to sexual selection [7], 59 of them focusing on questions associated with "pre-copulatory sexual traits" (i.e. sexual signals) [7]. In November 2023, PP selected 44 meta-analyses from this set, specifically the ones examining the relationship between sexual signals and distinct conditions, fitness proxies, or individual traits (hereby *proxies*; see Table S1). PP also included another meta-analysis examining the relationship between ornament expression and parasite load [14], which was published after searches were conducted in [7]. This resulted in a total of 45 eligible meta-analyses to fulfil our objectives [14–54,58–61].

PP gathered the data reported by these meta-analyses from their tables, appendices, supplementary files, and occasionally from direct correspondence with their authors. We could not obtain data for four eligible meta-analyses as their data were unreported [59–61] or poorly described [58]. Altogether, the remaining 41 meta-analyses from which PP extracted data [14–54] yielded 6,773 data points (see *Analyses* section).

PP filtered the collated dataset for most analyses conducted here (see *Analyses* section and Supplementary information S2). This was done because not all data collected from meta-analyses were relevant to objectives of other parts of the larger research project (e.g. meta-meta-analysis of sexual signals; see [55]). Following this filtering process, the number of data points in the collated dataset decreased to 5,496.

Re-extractions

Many meta-analyses included in the collated dataset extracted data from the same empirical studies. More specifically, after filtering, PP detected that a quarter of empirical studies in the dataset (318 out of 1272) were used as sources by at least two different meta-analyses (hereby *duplicates*, Figure S1). PP selected a subset of these duplicates for data re-extraction to reduce sampling effort (generating a greater sample size of data points for each meta-analysis assessed), prioritising certain empirical studies (see details in Supplementary information S3). In total, this process produced a set of 249 empirical studies for re-extraction. However, we could not access the full-text of three of these studies, so our sample of studies for re-extraction was reduced to 246 (hereby *verified* empirical studies; [62–307]).

We extracted all data on the relationship between sexual signals and proxies from verified empirical studies (see details in Supplementary information S2), blinded to which exact data points were extracted for each of the meta-analyses that included a given study. More specifically, PP extracted data from 59.3% of the selected empirical studies, while SN, YY, AM, RCMR, and ML respectively extracted data from other 15.5%, 7.7%, 6.5%, 5.7%, and 5.3% studies. PP then cross-checked all data extractions done by other authors (40.7% of studies). Importantly, authors never re-extracted data from empirical studies they participated in. However, we cannot rule out that we experienced some unconscious bias when dealing with the meta-analyses of co-authors. We extracted data required to calculate effect sizes from text, tables, supplementary material, and figures (using the package *metaDigitise* 1.0.1 [308]). When empirical studies reported similar results in various forms, we prioritised extractions in the following order: (1) raw data (calculating estimates directly) from sources other than figures, (2) raw estimates (i.e. means and correlation coefficients) from sources other than figures, (3) raw data or raw estimates from figures, (4) other estimates (e.g. t, β, χ^2) regardless of their origin. Nonetheless, we first prioritised data sources that showed more details. For instance, if a correlation was given for all individuals in-text but a scatter plot showed the same data with dots separated by age or sex, we collected data from the latter. In total, we extracted 5,703 valid data points. We note that we do not claim that our re-extracted data points are more or less correct than the data originally reported by meta-analyses, yet we do expect them to converge, meaning that mismatches should be taken seriously given our transparent procedures.

Matching reported data with re-extracted data

PP carefully examined the inclusion criteria reported in meta-analyses to verify which re-extracted data points should have been included by them. However, PP found several issues with these inclusion criteria. First, the proxies and sexual signals that meta-analyses included were often vague or ambiguous. For instance, Dougherty [15] was interested in behavioural sexual signals but their dataset also included extended phenotypes (e.g. domes built by crabs). Even though these extended phenotypes can be considered behavioural products, it was unclear whether other similar structures (e.g. bowers, ornamented nests) were deemed valid for inclusion by Dougherty [15]. Second, we detected apparent inclusion criteria patterns in datasets of some meta-analyses' that were not mentioned in-text. For example, Nakagawa et al.

[36] stated that they included data on reproductive success but the only reproductive success measure in their dataset was the number of fledglings, even though the studies they used for effect size extraction also contained other measures (e.g. number of eggs, number of hatchlings). Third, some meta-analyses outwardly contradicted their own information. For example, Weaver et al. [52] stated that they included standardised colour metrics (hue, chroma or composite measures of those) for carotenoid-based colours in adult birds, describing specific proxies in their table 2. Yet, they seemed to have included data points in which (1) the sexual signal was the size of a colourful plumage (e.g. patch size), (2) individuals were juveniles (including when data points were separated by age), and (3) proxies other than the ones reported in-text were used (e.g. offspring size). We summarised all meta-analyses' originally reported inclusion criteria, the ambiguities, omissions, and contradictions we detected in them, and how we dealt with these issues for matching purposes in Supplementary information S4.

We then attempted to match data points reported in meta-analyses with the ones we re-extracted from empirical sources. PP mainly used the description of sexual signals and proxies of each data point to match them with re-extracted data. When multiple data points from the same empirical study had similar descriptions, we also used sample size and other additional information (e.g. statistics reported, if given) for matching purposes.

There were three possibilities for each matching attempt. First, when both original and re-extracted data points had a similar description, PP linked them by labelling the latter with the ID of the former (i.e. successfully matched them). Yet, it was common to find multiple data points in our re-extracted dataset that matched one or many data points from the originally reported dataset (or vice-versa), so this matching was not necessarily exact (see examples in Supplementary information S5). Second, there were cases in which we could not find original data points with an equivalent description to relevant re-extracted ones. We assumed that these data were missed or undetected by meta-analyses' authors. This allowed us to obtain two aspects related to reliability: (1) the number of relevant data points that were absent in meta-analyses' datasets despite being present in empirical studies reported as sources (hereby *missing data*) and (2) the number of empirical studies that contained relevant data points that should have been used as sources but were not (hereby *undetected studies*). To clarify the latter, consider a hypothetical meta-analysis that investigated the relationship between X and Y, reporting data from 10 empirical papers. We then notice two other studies containing relevant data (relationship between X and Y) that were used as data sources by other meta-analyses in our dataset but not the hypothetical one, even though the data from these two empirical studies fit the inclusion criteria reported by the hypothetical meta-analysis. We thus deem that the hypothetical meta-analysis failed to detect at least two relevant articles. Third, there were cases in which data points were shown in meta-analyses' reported datasets but were absent in our re-extractions. We re-checked all of these latter cases: although some data points were not extracted by us because of our criteria (e.g. invalid proxy) or because we considered them repeated data, most of them could simply not be found in empirical studies (see details in Table S2). This could have happened if meta-analyses' authors contacted empirical authors and were thus able to obtain more data than what was shown in the empirical articles (unpublished data). Alternatively, meta-analyses' authors may have made mistakes during data extraction, even though we cannot ascertain when this was truly the case.

After PP finalised the matching process, AM, ML, RCMR, SN, and YY cross-checked matching decisions for five different empirical studies each. This resulted in matching decisions for approximately 10% of all verified studies being cross-checked, somewhat attesting the reliability of our process.

Effect sizes

Originally reported data points were given as the following effect size types: Cohen's *d* [18,20], logarithm of response ratio (logRR) [28], and Fisher's *Zr* or correlation coefficients (*r*) (remaining meta-analytical studies). One meta-analysis in particular [38] only provided *p*-values and sample sizes, so PP calculated effect sizes from this information. We transformed effect sizes reported to *Zr* for all analyses except the replication of general results (see *Analyses* section). We also calculated *Zr* (along with its sampling variance) from all re-extracted data points. Additionally, effect sizes reported by Koch et al. [28] did not reflect the raw data that they provided (means and standard deviations), so PP re-calculated their effect sizes. All equations for calculation and conversion of effect sizes are given in the Supplementary information S6, while the direction rationale applied to effect sizes is detailed in Supplementary information S7.

Analyses

First, we attempted to replicate meta-analyses' reported mean effect sizes using their original datasets. To do so, we conducted a meta-analytical model for each meta-analytical study using all of their reported effect sizes together (i.e. global model, *sensu* [7]). However, for meta-analytical studies that only performed subgroup analyses, we only re-analysed the largest or first reported subgroup. For instance, Nolazco et al. [37] only analysed the relationship between plumage colour and proxies for each sex separately, so we re-analysed only their data related to females. Effect sizes for these replication analyses were of the same type as results reported by meta-analyses (see *Effect sizes* section). Meta-analytical models for all of our analyses contained multiple random factors (see end of this section) but, for this replication analysis, we also included an additional random factor if provided by authors. For example, both Robinson & Creanza [42] and Sánchez-Tójar et al. [45] used population ID as an additional random factor in their models. Yet, we were unable to include these additional random factors in meta-analytic models when these variables were not provided by authors with the rest of the data (e.g. experiment ID in [15] and population ID in [46]) or when they were redundant (population ID was different for every source in [16]). Moreover, two meta-analytical studies from our collated dataset were excluded from this specific analysis: Parker et al. [38] did not provide a confidence interval for their mean effect size and only data points of interest were extracted from Thornhill et al. [51] (i.e. their dataset was not fully extracted). We also tested for signs of publication bias (small-study effect) in re-analysed datasets by adding the inverse of the effective sample size as a moderator in meta-analytical models (alternative Egger's regression) [309].

Second, we attempted to replicate part of individual effect sizes reported in meta-analyses. To do so, we compared originally reported effect sizes with the ones we re-extracted using linear models with only exactly matched data (i.e. reported data point matched to a single re-extracted data point, see Supplementary information S5). In these linear models, the intercept was forced to be zero, originally reported effect sizes were the response variable, and re-extracted effect sizes were the predictor variable. We evaluated whether the 95% confidence interval of the estimated slope in each linear model included the value 1, which would represent that reported and re-extracted effect sizes are very similar. We could not compare originally reported effect sizes with re-extracted effect sizes from 11 meta-analyses because they contained less than six exactly matched data points (the arbitrary threshold we established for this analysis), so we only reported results related to the remaining 30 meta-analyses in our dataset.

Third, we assessed meta-analyses' reproducibility by comparing the results of four slightly distinct meta-analytical models for each meta-analytical study. In model 1, we analysed the filtered dataset, which contained only reported effect sizes. In model 2, we used the same data as in model 1 but we replaced reported effect sizes that came from verified studies with their equivalent re-extracted effect sizes (i.e. the ones that matched in description). In model 3, we used the same data as in model 2 but we also added other relevant re-extracted data points from verified studies that were originally missed by meta-analyses' authors (i.e. missed data). In model 4, we used the same data as in model 3 but we also added all relevant effect sizes from undetected studies. Differently from our analysis attempting to replicate mean effect sizes (first described in this section), we conducted global models for all meta-analytical studies, even those that had done only subgroup analyses. We note that the datasets used for most meta-analyses remained similar across the four models described above because we did not re-extract data from all empirical studies reported as sources (Figure S2). For example, Dougherty [15] extracted data from 197 empirical studies, but we verified only 15 of these studies. This means that results from these distinct models were unlikely to change for meta-analyses with proportionally few verified sources. Thus, we additionally conducted the same four models described above using only data from verified studies (instead of using data from both verified and unverified studies) for meta-analyses with at least 15 verified sources (N = 15). Furthermore, we also evaluated whether the inclusion of missing data and undetected studies changed the generalisability of results by examining the σ (a measure of heterogeneity) from each random factor in meta-analytical models.

When applicable, we compared estimates from meta-analytic models both qualitatively and quantitatively. We first ascertained whether an estimate was positive, negative, or not different from zero (if its 95% CI overlapped zero), so that estimates with distinct classifications represented a qualitative difference. By contrast, a quantitative difference occurred when the absolute difference between two estimates produced a z-score of more than 1.96 (i.e. statistically significant, two-tailed $\alpha = 0.05$), calculated as:

$$z = \frac{X_{reported} - X_{re-analysed}}{\sqrt{se^2_{reported} + se^2_{re-analysed} - 2rse_{reported}se_{re-analysed}}}$$

where *X* represents the estimated mean effect size, *se* represents its standard error, and *r* represents the correlation coefficient between these two groups (set as 0.8 for all main results as we assume this correlation is high but not perfect; but see Figures S5, S6, and S7 for results using alternative methods). Although we showed all mean effect sizes generated in the last set of analyses, where we conducted up to four models per meta-analysis, we only discuss comparisons between the first and last models for simplicity.

We conducted all analyses described above in R 4.4.0 [310]. Multilevel meta-analytical models were fitted using the *rma.rv* function from the package *metafor* 4.6-0 [311]. All meta-analytical models fitted contained empirical study ID, species ID (non-phylogenetic effect), and phylogenetic relatedness as a random factor in meta-analytical models [312]. However, we removed phylogenetic relatedness from meta-analytical models related to certain articles [18,24,31,32], otherwise some of them would not converge. Phylogenetic trees were built using the packages *ape* 5.8 [313] and *rotl* 3.1.0 [314], which uses data from [315].

Results

Reproducibility of mean effect sizes

We found qualitative differences in 15.4% (6 out of 39) of the comparisons between mean effect sizes reported by meta-analyses and mean effect sizes resulting from re-analyses of those meta-analyses' reported datasets (Figure 1). More specifically, five of these meta-analyses reported a mean effect size not different from zero but its re-analysed counterpart was positive, while the opposite happened for another meta-analysis. We also detected quantitative differences between originally reported and re-analysed mean effect sizes in approximately a tenth of comparisons made (4 out of 39), although none of them were qualitative differences (Figure 1). Despite this, the absolute difference in magnitude between originally reported and re-analysed mean effect sizes was always lower than 0.2, even for cases with detected qualitative or quantitative differences (Figure 2). We also found that effect sizes were positively predicted by effective sample size (i.e. evidence of publication bias) in 23.1% (9 out of 39) of meta-analyses. Two-thirds of these meta-analyses (6 out of 9) diligently reported that they detected publication bias [23,24,26,27,39,49], while the remaining third did not assess publication bias whatsoever [17,20,54] (Table S3).

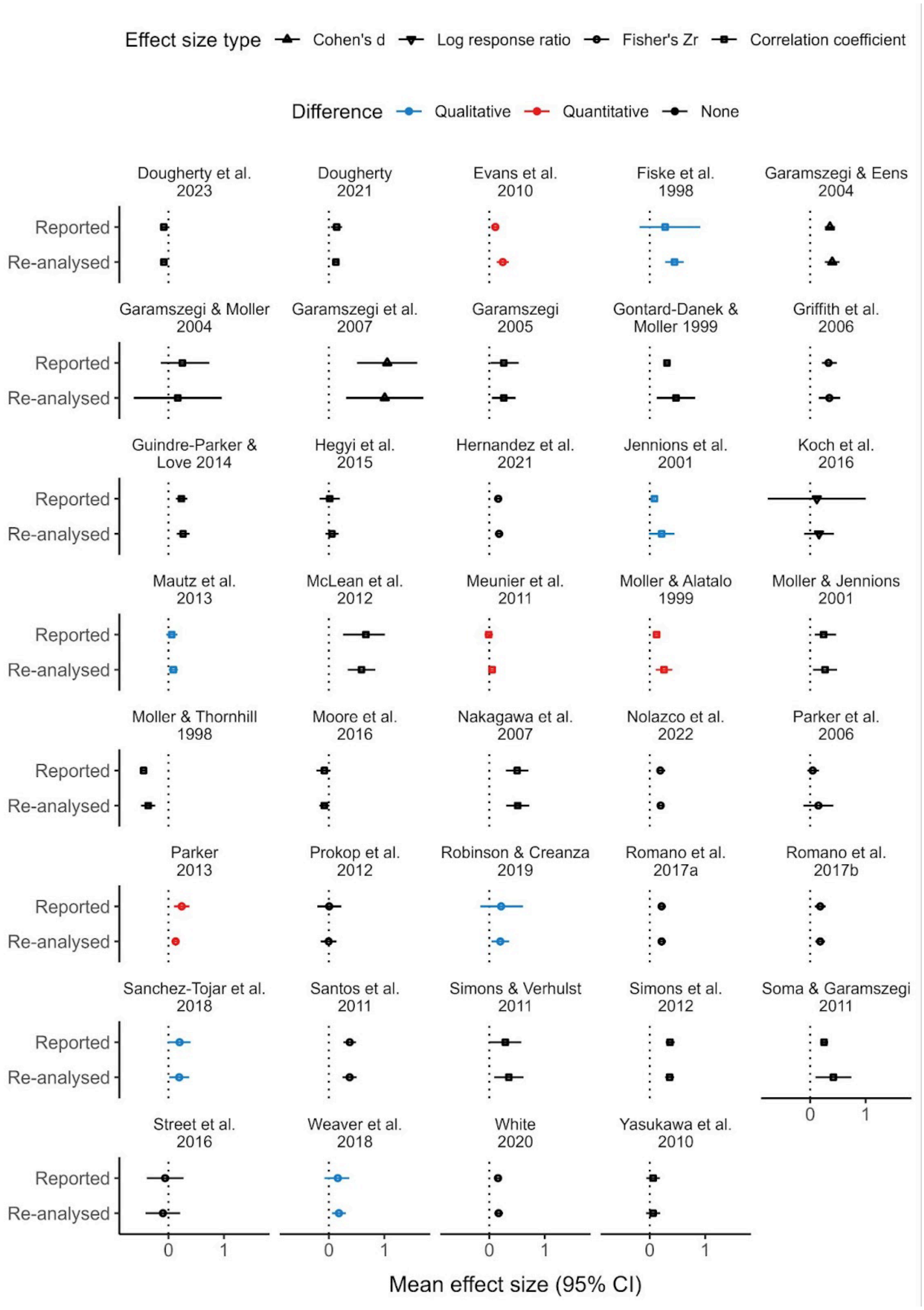

Figure 1. Originally reported and re-analysed mean effect sizes of 39 meta-analyses. Qualitative differences represent a change in interpretation between pairs (positive *vs*. not

different from zero or vice-versa), while quantitative differences represent statistical differences (absolute z-score greater than 1.96).

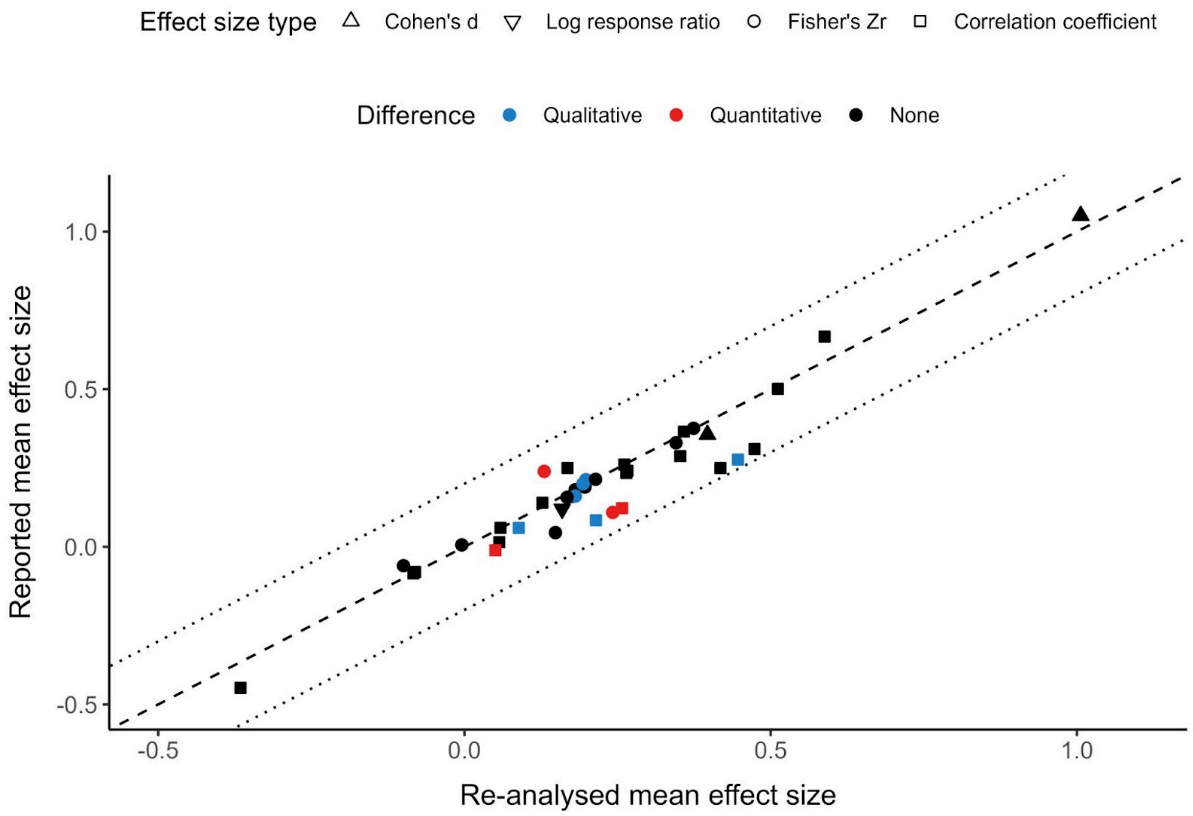

Figure 2. Comparison between mean effect sizes originally reported by meta-analyses and mean effect sizes resulting from the re-analysis of the reported dataset of the same meta-analyses. The dashed line highlights a perfect correspondence between variables, while dotted lines highlight a difference of 0.2 and -0.2 between reported and re-analysed mean effect sizes.

Reproducibility of individual effect sizes

We found that effect sizes originally reported by meta-analyses were, on average, statistically identical to their equivalent re-extracted ones in less than half (14 out of 30) of the meta-analyses evaluated (Figure 3, Table S4). For the remaining meta-analyses, the slope from linear regressions between reported and re-extracted effect sizes was always lower than 1 ($\bar{x}$±SE =

0.55±0.07), showing that reported effect sizes tended to be smaller or in the opposite direction than re-extracted ones in these cases.

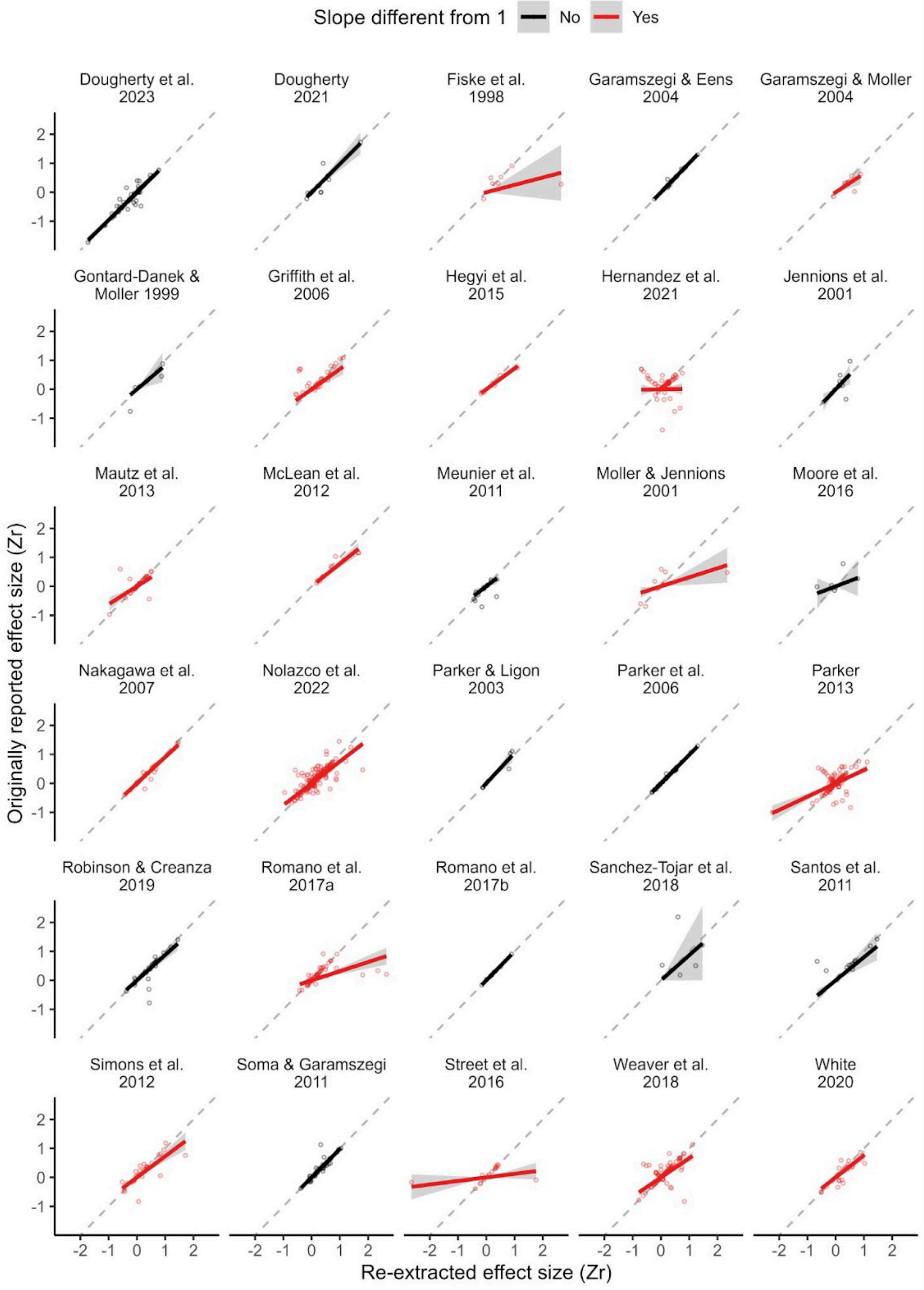

Figure 3. Relationship between effect sizes originally reported by meta-analyses and effect sizes re-extracted from empirical studies. Solid lines represent a linear regression between reported and re-extracted effect sizes (with the intercept forced to zero), while shaded areas represent their 95% confidence interval. Red lines indicate slopes that significantly differ from 1. Dashed lines highlight a perfect relationship between reported and re-extracted effect sizes.

Missing data and undetected studies

We found that meta-analyses missed (i.e. failed to extract and report), on average, 16±2.6% (x̄±SE) of relevant effect sizes from empirical studies they used as sources for data extraction (Figure 4A). Additionally, we found that meta-analyses were unsuccessful in detecting, on average, at least 10% of empirical studies they should have included as data sources (Figure 4B). Furthermore, we found no association between the number of missing effect sizes and the minimum proportion of undetected studies across meta-analyses (Figure 4C).

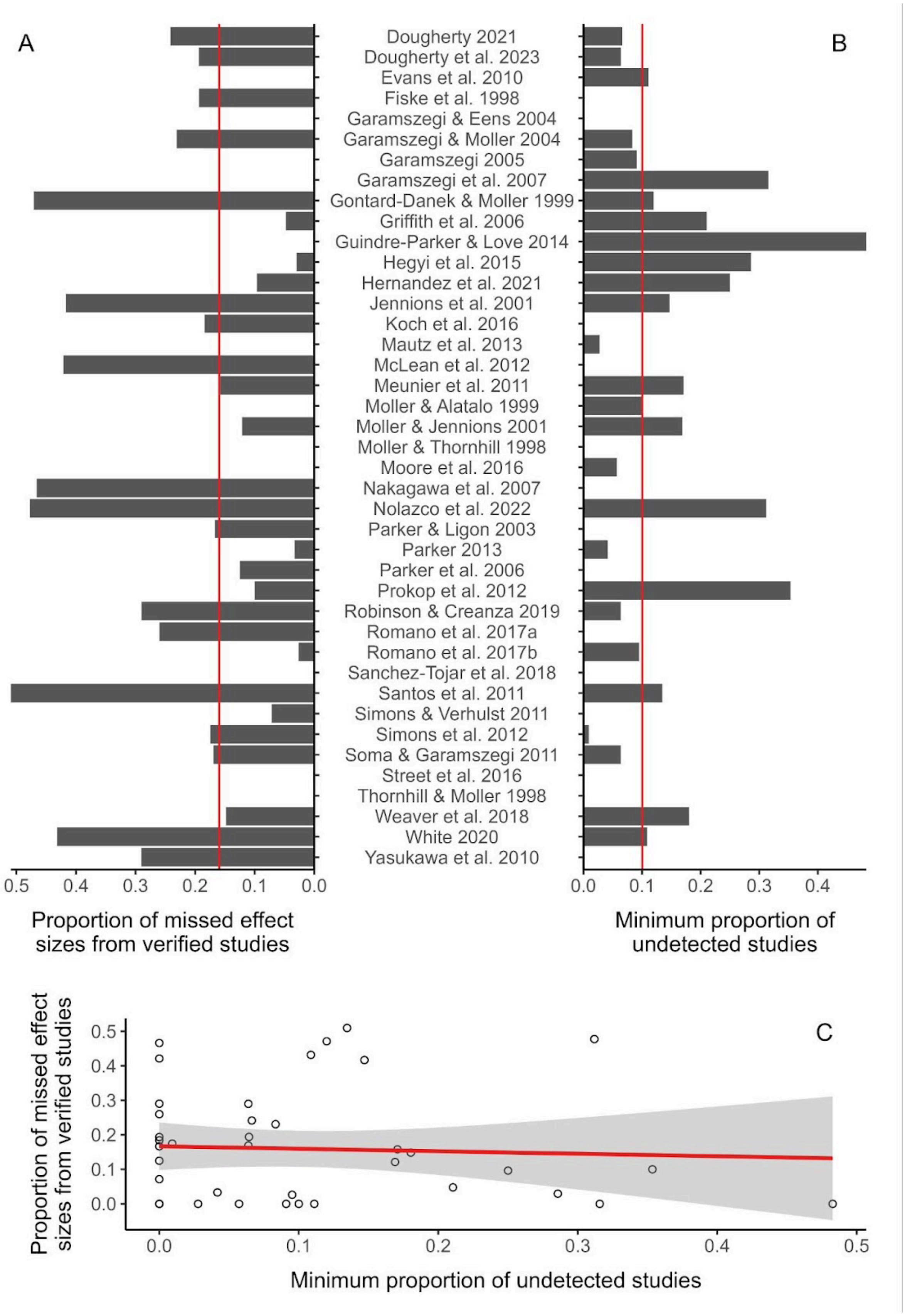

Figure 4. Proportion of missed effect sizes from all relevant effect sizes re-extracted from verified empirical studies (A), proportion of undetected studies from the minimum number of

empirical studies that each meta-analysis should have included (B), and the relationship between these variables (C). Vertical red lines in panels A and B represent the average proportion of missed effect sizes from all relevant re-extracted effect sizes and the minimum proportion of undetected studies, respectively. The red line in panel C represents the fit of linear regression between variables with its 95% confidence interval as the shaded area.

Replicability of results

Using data from both verified and unverified empirical studies, we found that mean effect sizes from analyses of originally reported datasets and those from analyses incorporating all relevant re-extracted data (matched, missed data, and undetected studies) were qualitative and quantitatively distinct for 12.2% (5 out of 41) and 9.7% (4 out of 41) of meta-analyses, respectively (both types of differences occurred for one meta-analysis; Figure 5, but see Figure S6). Similar comparisons using only data from verified empirical studies increased the occurrence of both qualitative differences (20%, 3 out of 15) and quantitative differences (26.7%, 4 out of 15; Figure 6, but see Figure S7). Nonetheless, differences in magnitude between mean effect sizes were often small (Figure S3). Furthermore, heterogeneity varied only slightly among fitted models (with perhaps one exception; Figure S4).

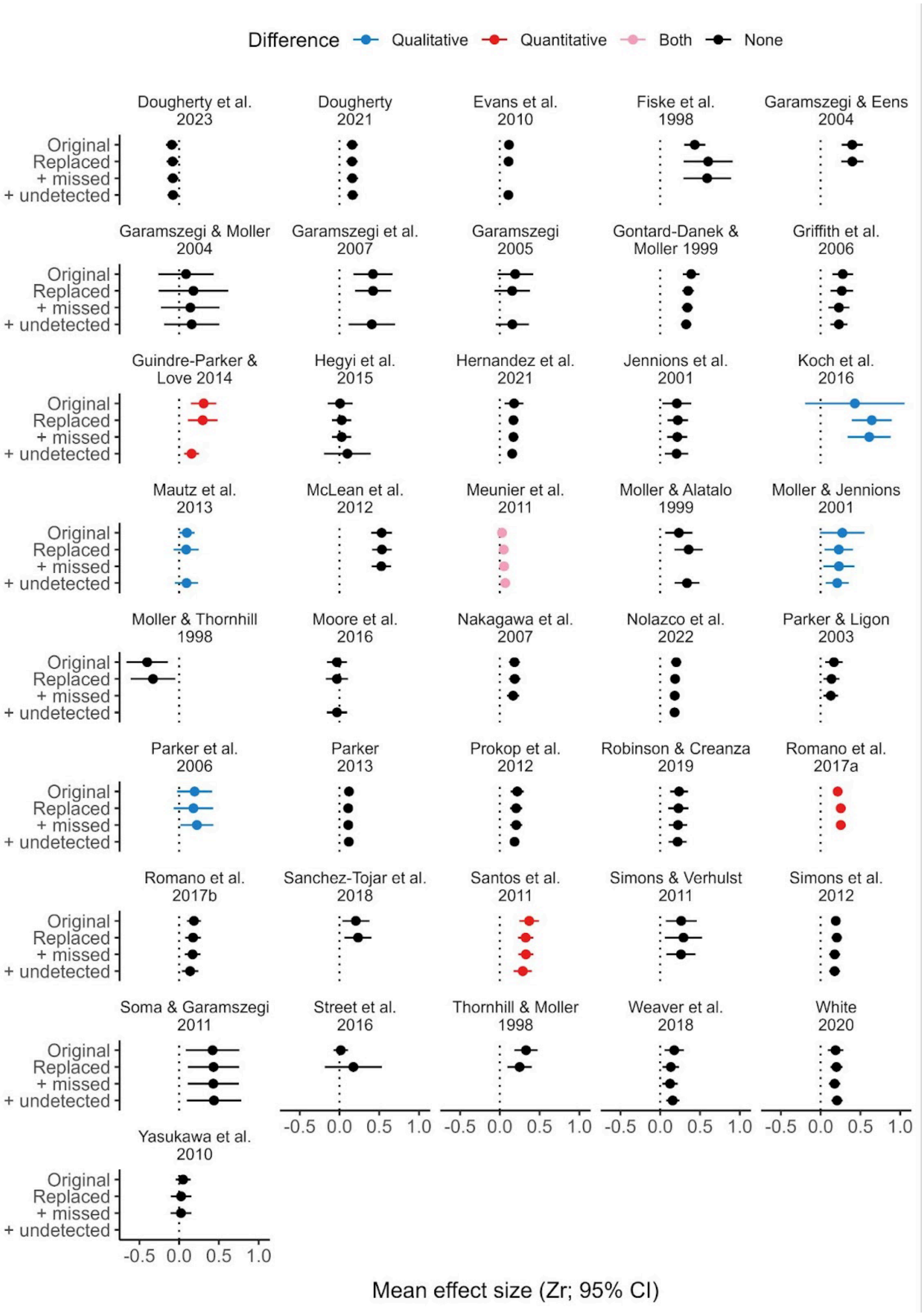

Figure 5. Mean effect sizes from up to four distinct meta-analytical models for each of 41 meta-analyses, using data from all empirical studies (both verified and unverified by us, see details

in-text). Dotted lines highlight zero. Comparisons were made between the first and the last result shown within each subplot, with qualitative differences representing a change in interpretation (positive *vs*. not different from zero or vice-versa) and quantitative differences representing statistical differences (absolute *z*-score greater than 1.96).

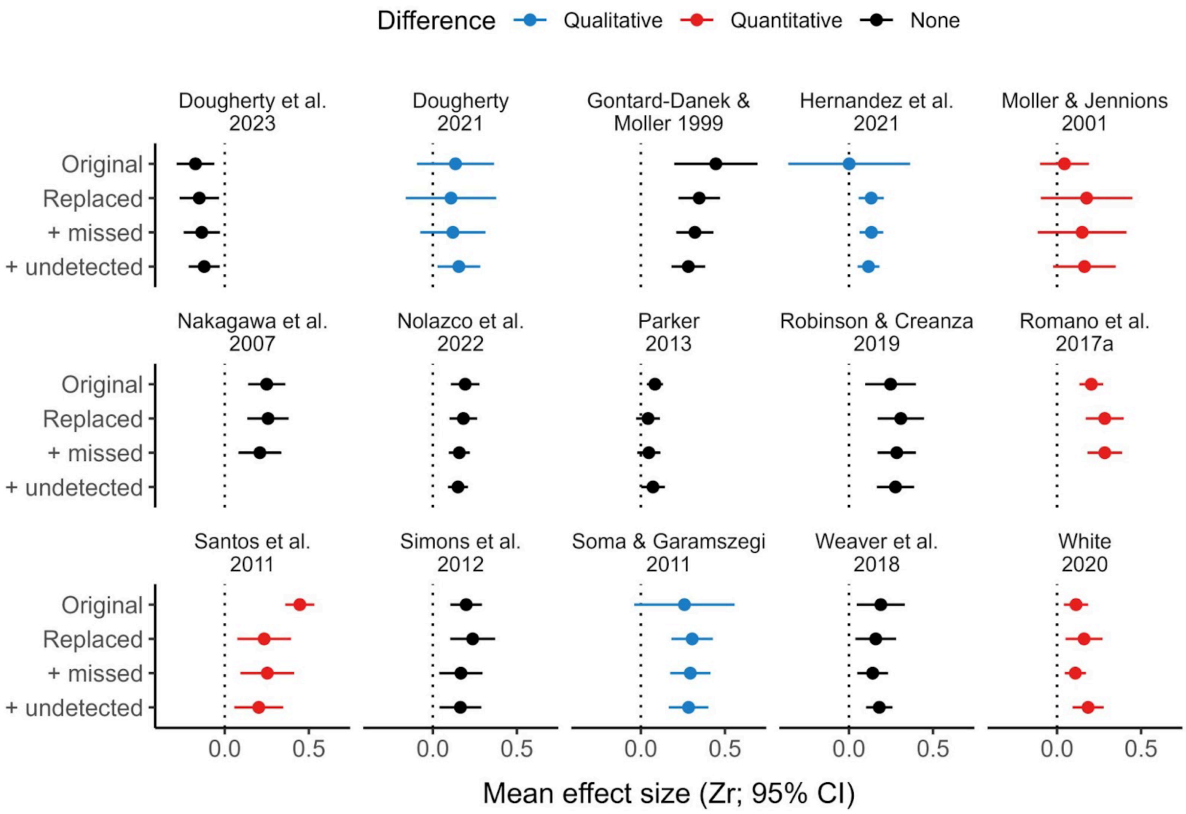

Figure 6. Mean effect sizes from up to four distinct meta-analytical models for each of 15 meta-analyses, using only data from empirical studies that we verified (see details in-text). Dotted lines highlight zero. Comparisons were made between the first and the last result shown within each subplot, with qualitative differences representing a change in interpretation (positive *vs*. not different from zero or vice-versa) and quantitative differences representing statistical differences (absolute *z*-score greater than 1.96).

Discussion

As the very first case study in ecology and evolution, we assessed multiple aspects related to the reliability of meta-analyses that focus on the relationship between sexual signals and various proxies. First, we were able to precisely replicate meta-analyses' reported mean effect sizes in most but not all cases. Second, we found that effect sizes reported by meta-analyses and those we re-extracted did not overlap in multiple cases. Third, we observed that some relevant data, both from included and undetected studies, were missing from certain meta-analyses. Fourth, incorporating such missed relevant data into further analyses, in addition to replacing reported data from verified empirical studies with re-extracted data, sometimes produced distinct meta-analytical results. However, the discrepancies we found tended to be small in magnitude and rarely generated distinct interpretations from the ones originally reported. Thus, we argue that the meta-analyses we evaluated are fairly reliable, although this is subjective to the degree of accuracy expected. Below we discuss the impacts of our findings and provide recommendations to further improve the reliability of meta-analyses in ecology and evolution.

Most of our results are based on comparisons between mean effect sizes (Figures 2, 5, and 6) using categorical interpretations (i.e. negative, not different from zero, or positive; indicating qualitative differences) and statistical calculations (indicating quantitative differences). We detected qualitative differences in only a handful of these comparisons, most of them resulting from differences in width of mean effect sizes' confidence intervals rather than in their value. Similarly, we detected few quantitative differences when comparing mean effect sizes that ideally would have remained identical across analyses. Still, these discrepancies are only relevant if they impact the interpretation of results from original meta-analyses. Ecology and evolution researchers rarely mention the magnitude of mean effect sizes from meta-analyses, focusing their attention on the existence or absence of effects or relationships (Pollo et al. *in prep.*). This means that quantitative differences alone would probably be unimportant for most researchers (unless hypotheses rely on effect size comparisons, as in [316]), while qualitative differences would affect how results are perceived. However, even though interpreting results based on the inclusion of zero by confidence intervals is a common practice, it can generate misleading conclusions [317]. For instance, the lower confidence bound for the mean effect size of Sánchez-Tójar et al. [45] was reported as -0.01 but was slightly above zero in all our analyses, while the magnitude of this study's mean effect size was reported as 0.2 and ranged from 0.19 to 0.21 in our analyses (Figures 2 and 5). Therefore, we conclude that results from Sánchez-Tójar et al. [45], as well those from meta-analyses with similar minute discrepancies, should be deemed replicable despite being different to the ones we obtained when comparing the inclusion of zero by confidence intervals. Importantly, statistical definitions of replicability and reproducibility are highly debatable, meaning that conclusions based on result comparisons become inevitably subjective to a certain degree [318,319].

Some effect sizes reported in meta-analyses were very different from the ones we re-extracted (Figure 3). A portion of these discrepancies could be explained by using distinct effect size calculations and/or from extracting data from distinct sources within empirical studies (e.g. raw data from a figure *vs*. statistics reported in-text). Yet, we noticed that numerous effect sizes reported were similar in value but opposite in direction to the ones we re-extracted. Although this did not strongly affect differences among estimated mean effect sizes, effect sizes with the wrong direction can be particularly dangerous as they should be more impactful than simply imprecise ones, at least for large effect sizes. Furthermore, errors in data extraction or effect size calculation can especially affect meta-regressions, which are often performed with subsets of meta-analytical datasets. This emphasises the importance in meta-analyses of (1) establishing a coherent rationale to ascertain direction of effect sizes, (2) reporting directionality decisions in detail, and (3) remaining vigilant during data extraction.

Additionally, we recommend cross-checking data extractions (i.e. independent verification of the data by someone that did not extract them) to increase the chances of spotting and correcting mistakes, including those related to the direction of effect sizes.

We observed that not all relevant data from verified empirical studies were used by meta-analyses (Figure 4). Although we cannot ascertain how or why meta-analyses' authors missed relevant data from studies they extracted data from, the reason that some empirical studies remained undetected might be more easily explained. While it is possible that meta-analyses could have employed suboptimal searches, empirical studies commonly neglect to inform their entire scope and results in their title, abstract, and key words, which are used to retrieve and screen studies [320]. Thus, even though authors of meta-analyses should follow certain guidelines to build effective searches (e.g. [321]), meta-analysts may unfortunately miss relevant studies despite their best efforts. Conversely, authors of empirical studies should be mindful of how search engines work, crafting their title, abstract, and keywords to enhance the findability of their work [320,322].

The goal of our study was to evaluate the reliability of meta-analyses beyond transparency, yet our results might be affected by transparency issues found in the meta-analyses we assessed. First, matching reported and re-extracted data points proved to be a difficult task because meta-analyses usually lack details on their extracted data. For instance, the sexual signal and proxy for each data point was only vaguely described in most meta-analyses. Furthermore, no meta-analysis in our dataset reported the location of the extracted data within empirical studies (e.g. which page, table, figure, etc). Second, meta-analyses were often poorly transparent with their inclusion criteria. We tried to comprehend vague information, incorporate omitted criteria, and ignore contradictions when matching re-extracted to reported datasets (see Supplementary information S4). Nonetheless, our decisions might have affected the amount of missing data and of undetected studies by each meta-analysis. Ultimately, this could have influenced our findings related to the reproducibility of mean effect sizes.

In addition to the recommendations we already mentioned (e.g. cross-checking of data extracted), we urge meta-analysts to provide all possible details on the data they collect. For instance, mentioning which exact measurements were sought are necessary instead of simply mentioning umbrella terms (e.g. condition-dependence by [24]). The location of the information is also crucial: readers should not have to examine datasets to find important details. Instead, summarised details should be in the manuscript or, less preferably, in the supplementary material, but always in a readable format (figures, simplified tables, in-text, not in spreadsheets). Furthermore, we reiterate recommendations by Ivimey-Cook et al. [5], such as providing the within-text source for each data point extracted and the equations used to calculate effect sizes (along with assumptions and transformations utilised). We summarise our recommendations in Table S5.

Data and code availability

All data and code used in this study are available at: https://osf.io/6njem/?view_only=7b01538fb32e4f78b7130b6e8f303649.

Author contributions

Conceptualisation: PP, ML, AM, YY, SN; data curation: PP; formal analysis: PP; funding acquisition: SN; investigation: PP, ML, RCMR, AM, YY, SN; methodology: PP, ML, RCMR, AM, YY, SN; project administration: PP; software: PP; supervision: SN; visualisation: PP; writing – original draft: PP, SN; writing – review & editing: PP, ML, RCMR, AM, YY, SN.

Acknowledgements

We thank László Zsolt Garamszegi, Sarah Guindre-Parker, América Hernández Martínez, and Tim Parker for sending us data from their meta-analyses. We also thank Yong Zhi Foo for contributing to the conceptualisation of this project and Alfredo Sánchez-Tójar for feedback on the manuscript. PP, ML, YY, and SN were supported by ARC (Australian Research Council) Discovery Project grants (DP210100812 and DP230101248). SN was also supported by the Canada Excellence Research Chair Program (CERC-2022-00074). RCMR was supported by a grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) as part of the Programa de Desenvolvimento da Pós-Graduação (PDPG 88887.803974/2023-00).

Competing interests

We declare no competing interests.

References

1. McCallen E, Knott J, Nunez-Mir G, Taylor B, Jo I, Fei S. 2019 Trends in ecology: shifts in ecological research themes over the past four decades. *Front Ecol Environ* **17**, 109–116. (doi:10.1002/fee.1993)
2. Nunez-Mir GC, Iannone B V., Pijanowski BC, Kong N, Fei S. 2016 Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods Ecol Evol* **7**, 1262–1272. (doi:10.1111/2041-210X.12602)
3. Koricheva J, Gurevitch J, Mengersen K. 2013 *Handbook of meta-analysis in ecology and evolution*. Princeton, NJ: Princeton University Press.
4. Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. 2021 *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons.

5. Ivimey-Cook ER, Noble DWA, Nakagawa S, Lajeunesse MJ, Pick JL. 2023 Advice for improving the reproducibility of data extraction in meta-analysis. *Res Synth Methods* **14**, 911–915. (doi:10.1002/jrsm.1663)

6. O'Dea RE *et al.* 2021 Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension. *Biol Rev* **96**, 1695–1722. (doi:10.1111/brv.12721)

7. Pollo P, Lagisz M, Yang Y, Culina A, Nakagawa S. 2024 Synthesis of sexual selection: a systematic map of meta-analyses with bibliometric analysis. *Biol Rev* **4**. (doi:10.1111/brv.13117)

8. Stroup DF *et al.* 2000 Meta-analysis of observational studies in epidemiology. *JAMA* **283**, 2008. (doi:10.1001/jama.283.15.2008)

9. Page MJ *et al.* 2021 The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71. (doi:10.1136/bmj.n71)

10. Kambouris S, Wilkinson DP, Smith ET, Fidler F. 2024 Computationally reproducing results from meta-analyses in ecology and evolutionary biology using shared code and data. *PLoS One* **19**, e0300333. (doi:10.1371/journal.pone.0300333)

11. Stubblefield JW, Orzack SH. 2013 Resource transfers and evolution: helpful offspring and sex allocation. *Theor Popul Biol* **83**, 64–81. (doi:10.1016/j.tpb.2012.09.004)

12. Page MJ *et al.* 2021 The REPRISE project: protocol for an evaluation of REProducibility and Replicability In Syntheses of Evidence. *Syst Rev* **10**, 112. (doi:10.1186/s13643-021-01670-0)

13. Maassen E, Van Assen MALM, Nuijten MB, Olsson-Collentine A, Wicherts JM. 2020 Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One* **15**, 1–18. (doi:10.1371/journal.pone.0233107)

14. Dougherty LR, Rovenolt F, Luyet A, Jokela J, Stephenson JF. 2023 Ornaments indicate parasite load only if they are dynamic or parasites are contagious. *Evol Lett* **7**, 176–190. (doi:10.1093/evlett/qrad017)

15. Dougherty LR. 2021 Meta-analysis reveals that animal sexual signalling behaviour is honest and resource based. *Nat Ecol Evol* **5**, 688–699. (doi:10.1038/s41559-021-01409-z)

16. Evans SR, Hinks AE, Wilkin TA, Sheldon BC. 2010 Age, sex and beauty: methodological dependence of age- and sex-dichromatism in the great tit *Parus major*. *Biol J Linn Soc* **101**, 777–796. (doi:10.1111/j.1095-8312.2010.01548.x)

17. Fiske P, Rintamaki PT, Karvonen E. 1998 Mating success in lekking males: a meta-analysis. *Behav Ecol* **9**, 328–338. (doi:10.1093/beheco/9.4.328)

18. Garamszegi LZ, Eens M. 2004 Brain space for a learned task: strong intraspecific evidence for neural correlates of singing behavior in songbirds. *Brain Res Rev* **44**, 187–193. (doi:10.1016/j.brainresrev.2003.12.001)

19. Garamszegi LZ, Møller AP. 2004 Extrapair paternity and the evolution of bird song. *Behav Ecol* **15**, 508–519. (doi:10.1093/beheco/arh041)

20. Garamszegi LZ, Török J, Hegyi G, Szöllõsi E, Rosivall B, Eens M. 2007 Age-dependent expression of song in the collared flycatcher, *Ficedula albicollis*. *Ethology* **113**, 246–256. (doi:10.1111/j.1439-0310.2007.01337.x)

21. Garamszegi LZ. 2005 Bird song and parasites. *Behav Ecol Sociobiol* **59**, 167–180. (doi:10.1007/s00265-005-0041-8)

22. Gontard-Danek M-C, Møller AP. 1999 The strength of sexual selection: a meta-analysis of bird studies. *Behav Ecol* **10**, 476–486. (doi:10.1093/beheco/10.5.476)

23. Griffith SC, Parker TH, Olson VA. 2006 Melanin- versus carotenoid-based sexual signals: is the difference really so black and red? *Anim Behav* **71**, 749–763. (doi:10.1016/j.anbehav.2005.07.016)

24. Guindre-Parker S, Love OP. 2014 Revisiting the condition-dependence of melanin-based plumage. *J Avian Biol* **45**, 29–33. (doi:10.1111/j.1600-048X.2013.00190.x)

25. Hegyi G, Kötél D, Laczi M. 2015 Direct benefits of mate choice: a meta-analysis of plumage colour and offspring feeding rates in birds. *Sci Nat* **102**, 62. (doi:10.1007/s00114-015-1311-6)

26. Hernández A, Martínez-Gómez M, Beamonte-Barrientos R, Montoya B. 2021 Colourful traits in female birds relate to individual condition, reproductive performance and male-mate preferences: a meta-analytic approach. *Biol Lett* **17**, 20210283. (doi:10.1098/rsbl.2021.0283)

27. Jennions MD, Møller AP, Petrie M. 2001 Sexually selected traits and adult survival: a meta-analysis. *Q Rev Biol* **76**, 3–36. (doi:10.1086/393743)

28. Koch RE, Wilson AE, Hill GE. 2016 The importance of carotenoid dose in supplementation studies with songbirds. *Physiol Biochem Zool* **89**, 61–71. (doi:10.1086/684485)

29. Mautz BS, Møller AP, Jennions MD. 2013 Do male secondary sexual characters signal ejaculate quality? A meta-analysis. *Biol Rev* **88**, 669–682. (doi:10.1111/brv.12022)

30. McLean MJ, Bishop PJ, Nakagawa S. 2012 Male quality, signal reliability and female choice: assessing the expectations of inter-sexual selection. *J Evol Biol* **25**, 1513–1520. (doi:10.1111/j.1420-9101.2012.02533.x)

31. Meunier J, Pinto SF, Burri R, Roulin A. 2011 Eumelanin-based coloration and fitness parameters in birds: a meta-analysis. *Behav Ecol Sociobiol* **65**, 559–567. (doi:10.1007/s00265-010-1092-z)

32. Møller AP, Alatalo R V. 1999 Good-genes effects in sexual selection. *Proc R Soc B* **266**, 85–91. (doi:10.1098/rspb.1999.0607)

33. Møller AP, Jennions M. 2001 How important are direct fitness benefits of sexual selection? *Naturwissenschaften* **88**, 401–415. (doi:10.1007/s001140100255)

34. Møller AP, Thornhill R. 1998 Bilateral symmetry and sexual selection: a meta-analysis. *Am Nat* **151**, 174–192. (doi:10.1086/286110)

35. Moore FR, Shuker DM, Dougherty L. 2016 Stress and sexual signaling: a systematic review and meta-analysis. *Behav Ecol* **27**, 363–371. (doi:10.1093/beheco/arv195)

36. Nakagawa S, Ockendon N, Gillespie DOS, Hatchwell BJ, Burke T. 2007 Assessing the function of house sparrows' bib size using a flexible meta-analysis method. *Behav Ecol* **18**, 831–840. (doi:10.1093/beheco/arm050)

37. Nolazco S, Delhey K, Nakagawa S, Peters A. 2022 Ornaments are equally informative in male and female birds. *Nat Commun* **13**, 5917. (doi:10.1038/s41467-022-33548-7)

38. Parker TH, Ligon JD. 2003 Female mating preferences in red junglefowl: a meta-analysis. *Ethol Ecol Evol* **15**, 63–72. (doi:10.1080/08927014.2003.9522691)

39. Parker TH. 2013 What do we really know about the signalling role of plumage colour in blue tits? A case study of impediments to progress in evolutionary biology. *Biol Rev* **88**, 511–536. (doi:10.1111/brv.12013)

40. Parker TH, Barr IR, Griffith SC. 2006 The blue tit's song is an inconsistent signal of male condition. *Behav Ecol* **17**, 1029–1040. (doi:10.1093/beheco/arl041)

41. Prokop ZM, Michalczyk Ł, Drobniak SM, Herdegen M, Radwan J. 2012 Meta-analysis suggests choosy females get sexy sons more than 'good genes'. *Evolution* **66**, 2665–2673. (doi:10.1111/j.1558-5646.2012.01654.x)

42. Robinson CM, Creanza N. 2019 Species-level repertoire size predicts a correlation between individual song elaboration and reproductive success. *Ecol Evol* **9**, 8362–8377. (doi:10.1002/ece3.5418)

43. Romano A, Costanzo A, Rubolini D, Saino N, Møller AP. 2017 Geographical and seasonal variation in the intensity of sexual selection in the barn swallow *Hirundo rustica*: a meta-analysis. *Biol Rev* **92**, 1582–1600. (doi:10.1111/brv.12297)

44. Romano A, Saino N, Møller AP. 2017 Viability and expression of sexual ornaments in the barn swallow *Hirundo rustica*: a meta-analysis. *J Evol Biol* **30**, 1929–1935. (doi:10.1111/jeb.13151)

45. Sánchez-Tójar A *et al.* 2018 Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. *eLife* **7**, 1–26. (doi:10.7554/eLife.37385)

46. Santos ESA, Scheck D, Nakagawa S. 2011 Dominance and plumage traits: meta-analysis and metaregression analysis. *Anim Behav* **82**, 3–19. (doi:10.1016/j.anbehav.2011.03.022)

47. Simons MJP, Verhulst S. 2011 Zebra finch females prefer males with redder bills independent of song rate–a meta-analysis. *Behav Ecol* **22**, 755–762. (doi:10.1093/beheco/arr043)

48. Simons MJP, Cohen AA, Verhulst S. 2012 What does carotenoid-dependent coloration tell? Plasma carotenoid level signals immunocompetence and oxidative stress state in birds-a meta-analysis. *PLoS One* **7**, e43088. (doi:10.1371/journal.pone.0043088)

49. Soma M, Garamszegi LZ. 2011 Rethinking birdsong evolution: meta-analysis of the relationship between song complexity and reproductive success. *Behav Ecol* **22**, 363–371. (doi:10.1093/beheco/arq219)

50. Street SE, Cross CP, Brown GR. 2016 Exaggerated sexual swellings in female nonhuman primates are reliable signals of female fertility and body condition. *Anim Behav* **112**, 203–212. (doi:10.1016/j.anbehav.2015.11.023)

51. Thornhill R, Møller AP. 1998 The relative importance of size and asymmetry in sexual selection. *Behav Ecol* **9**, 546–551. (doi:10.1093/beheco/9.6.546)

52. Weaver RJ, Santos ESA, Tucker AM, Wilson AE, Hill GE. 2018 Carotenoid metabolism strengthens the link between feather coloration and individual quality. *Nat Commun* **9**, 73. (doi:10.1038/s41467-017-02649-z)

53. White TE. 2020 Structural colours reflect individual quality: a meta-analysis. *Biol Lett* **16**, 20200001. (doi:10.1098/rsbl.2020.0001)

54. Yasukawa K, Enstrom DA, Parker PG, Jones TC. 2010 Male Red-winged Blackbirds with experimentally dulled epaulets experience no disadvantage in sexual selection. *J Field Ornithol* **81**, 31–41. (doi:10.1111/j.1557-9263.2009.00258.x)

55. Pollo P, Foo YZ, Lagisz M, Macedo-Rego RC, Mizuno A, Yang Y, Nakagawa S. 2024 Pre-registration for a methodological evaluation and a meta-meta-analysis of sexual signals. (doi:10.17605/OSF.IO/HR2YK)

56. Nakagawa S *et al.* 2023 Method Reporting with Initials for Transparency (MeRIT) promotes more granularity and accountability for author contributions. *Nat Commun* **14**, 1788. (doi:10.1038/s41467-023-37039-1)

57. McNutt MK *et al.* 2018 Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proc Natl Acad Sci U S A* **115**, 2557–2560. (doi:10.1073/pnas.1715374115)

58. Møller AP, Christe P, Lux E. 1999 Parasitism, host immune function, and sexual selection. *Q Rev Biol* **74**, 3–20. (doi:10.1086/392949)

59. Hamilton WJ, Poulin R. 1997 The Hamilton and Zuk hypothesis revisited: a meta-analytical approach. *Behaviour* **134**, 299–320. (doi:10.1163/156853997X00485)

60. Leung B, Forbes MR. 1996 Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. *Écoscience* **3**, 400–413. (doi:10.1080/11956860.1996.11682357)

61. Macedo-Rego RC, Jennions MD, Santos ESA. 2020 Larger females and larger males produce more offspring, regardless of the social mating system: a meta-analysis on animals. In *When and how do episodes of sexual selection occur in animal species? A meta-analytical investigation* (ed RC Macedo-Rego), pp. 148–187. São Paulo, Brazil: Universidade de São Paulo.

62. Postma E, Gienapp P. 2009 Origin-related differences in plumage coloration within an island population of great tits (*Parus major*). *Can J Zool* **87**, 1–7. (doi:10.1139/Z08-132)

63. Gil D, Slater PJB. 2000 Multiple song repertoire characteristics in the willow warbler (*Phylloscopus trochilus*): correlations with female choice and offspring viability. *Behav Ecol Sociobiol* **47**, 319–326. (doi:10.1007/s002650050672)

64. Howard RD. 1974 The influence of sexual selection and interspecific competition on mockingbird song (*Mimus polyglottos*). *Evolution* **28**, 428–438. (doi:10.1111/j.1558-5646.1974.tb00764.x)

65. Horn AG, Dickinson TE, Falls JB. 1993 Male quality and song repertoires in western meadowlarks (*Sturnella neglecta*). *Can J Zool* **71**, 1059–1061. (doi:10.1139/z93-140)

66. Ferns PN, Hinsley SA. 2008 Carotenoid plumage hue and chroma signal different aspects of individual and habitat quality in tits. *Ibis* **150**, 152–159. (doi:10.1111/j.1474-919X.2007.00759.x)

67. Searcy WA. 1984 Song repertoire size and female preferences in song sparrows. *Behav Ecol Sociobiol* **14**, 281–286. (doi:10.1007/BF00299499)

68. Forstmeier W, Kempenaers B, Meyer A, Leisler B. 2002 A novel song parameter correlates with extra-pair paternity and reflects male longevity. *Proc R Soc B* **269**, 1479–1485. (doi:10.1098/rspb.2002.2039)

69. Hiebert SM, Stoddard PK, Arcese P. 1989 Repertoire size, territory acquisition and reproductive success in the song sparrow. *Anim Behav* **37**, 266–273. (doi:10.1016/0003-3472(89)90115-2)

70. Lampe HM, Espmark YO. 2002 Mate choice in pied flycatchers *Ficedula hypoleuca*: can females use song to find high-quality males and territories? *Ibis* **145**, E24–E33. (doi:10.1046/j.1474-919x.2003.00144.x)

71. Bouwman KM, van Dijk RE, Wijmenga JJ, Komdeur J. 2007 Older male reed buntings are more successful at gaining extrapair fertilizations. *Anim Behav* **73**, 15–27. (doi:10.1016/j.anbehav.2006.01.031)

72. Hansen AJ, Rohwer S. 1986 Coverable badges and resource defence in birds. *Anim Behav* **34**, 69–76. (doi:10.1016/0003-3472(86)90007-2)

73. Røskaft E, Rohwer S. 1987 An experimental study of the function of the red epaulettes and the black body colour of male red-winged blackbirds. *Anim Behav* **35**, 1070–1077. (doi:10.1016/S0003-3472(87)80164-1)

74. Yasukawa K, Butler LK, Enstrom DA. 2009 Intersexual and intrasexual consequences of epaulet colour in male red-winged blackbirds: an experimental approach. *Anim Behav* **77**, 531–540. (doi:10.1016/j.anbehav.2008.11.013)

75. Aweida MK. 1995 Repertoires, territory size and mate attraction in western meadowlarks. *Condor* **97**, 1080–1083. (doi:10.2307/1369552)

76. Mora AR, Meniri M, Glauser G, Vallat A, Helfenstein F. 2016 Badge size reflects sperm oxidative status within social groups in the house sparrow *Passer domesticus*. *Front Ecol Evol* **4**, 1–10. (doi:10.3389/fevo.2016.00067)

77. Garamszegi LZ, Heylen D, Møller AP, Eens M, De Lope F. 2005 Age-dependent health status and song characteristics in the barn swallow. *Behav Ecol* **16**, 580–591. (doi:10.1093/beheco/ari029)

78. Balsby TJS. 2000 Song activity and variability in relation to male quality and female choice in whitethroats *Sylvia communis*. *J Avian Biol* **31**, 56–62. (doi:10.1034/j.1600-048X.2000.310108.x)

79. Eens M, Pinxten R, Verheyen RF. 1991 Male song as a cue for mate choice in the European starling. *Behaviour* **116**, 210–238. (doi:10.1163/156853991X00049)

80. Dreiss A, Richard M, Moyen F, White J, Møller AP, Danchin E. 2006 Sex ratio and male sexual characters in a population of blue tits, *Parus caeruleus*. *Behav Ecol* **17**, 13–19. (doi:10.1093/beheco/ari089)

81. Foerster K, Poesel A, Kunc H, Kempenaers B. 2002 The natural plasma testosterone profile of male blue tits during the breeding season and its relation to song output. *J Avian Biol* **33**, 269–275. (doi:10.1034/j.1600-048X.2002.330309.x)

82. Poesel A, Foerster K, Kempenaers B. 2001 The dawn song of the blue tit *Parus caeruleus* and its role in sexual selection. *Ethology* **107**, 521–531. (doi:10.1046/j.1439-0310.2001.00680.x)

83. Doutrelant C, Blondel J, Perret P, Lambrechts MM. 2000 Blue tit song repertoire size, male quality and interspecific. *J Avian Biol* **31**, 360–366. (doi:10.1034/j.1600-048X.2000.310312.x)

84. Piersma T, Mendes L, Hennekens J, Ratiarison S, Groenewold S, Jukema J. 2001 Breeding plumage honestly signals likelihood of tapeworm infestation in females of a long-distance migrating shorebird, the bar-tailed godwit. *Zoology* **104**, 41–48. (doi:10.1078/0944-2006-00003)

85. Conner RN, Anderson ME, Dickson JG. 1986 Relationships among territory size, habitat, song, and nesting success of northern cardinals. *Auk* **103**, 23–31. (doi:10.1093/auk/103.1.23)

86. Bijnens L. 1988 Blue tit parus caeruleus song in relation to survival, reproduction and biometry. *Bird Study* **35**, 61–67. (doi:10.1080/00063658809476980)

87. Freeman-Gallant CR, Amidon J, Berdy B, Wein S, Taff CC, Haussmann MF. 2011 Oxidative damage to DNA related to survivorship and carotenoid-based sexual ornamentation in the common yellowthroat. *Biol Lett* **7**, 429–432. (doi:10.1098/rsbl.2010.1186)

88. Freeman-Gallant CR, Taff CC, Morin DF, Dunn PO, Whittingham LA, Tsang SM. 2010 Sexual selection, multiple male ornaments, and age- and condition-dependent signaling in the common yellowthroat. *Evolution* **64**, 1007–1017. (doi:10.1111/j.1558-5646.2009.00873.x)

89. Byers BE. 2007 Extrapair paternity in chestnut-sided warblers is correlated with consistent vocal performance. *Behav Ecol* **18**, 130–136. (doi:10.1093/beheco/arl058)

90. Rehsteiner U, Geisser H, Reyer HU. 1998 Singing and mating success in water pipits: one specific song element makes all the difference. *Anim Behav* **55**, 1471–1481. (doi:10.1006/anbe.1998.0733)

91. Marshall RC, Buchanan KL, Catchpole CK. 2007 Song and female choice for extrapair copulations in the sedge warbler, *Acrocephalus schoenobaenus*. *Anim Behav* **73**, 629–635. (doi:10.1016/j.anbehav.2006.06.011)

92. Buchanan KL, Catchpole CK. 1997 Female choice in the sedge warbler, *Acrocephalus schoenobaenus*: multiple cues from song and territory quality. *Proc R Soc B* **264**, 521–526. (doi:10.1098/rspb.1997.0074)

93. Bell BD, Borowiec M, Lontkowski J, Pledger S. 2004 Short records of marsh warbler (*Acrocephalus palustris*) song provide indices that correlate with nesting success. *J Ornithol* **145**, 8–15. (doi:10.1007/s10336-003-0001-x)

94. Hadfield JD, Burgess MD, Lord A, Phillimore AB, Clegg SM, Owens IPF. 2006 Direct versus indirect sexual selection: Genetic basis of colour, size and recruitment in a wild bird. *Proc R Soc B* **273**, 1347–1353. (doi:10.1098/rspb.2005.3459)

95. Catchpole CK. 1986 Song repertoires and reproductive success in the great reed warbler *Acrocephalus arundinaceus*. *Behav Ecol Sociobiol* **19**, 439–445. (doi:10.1007/BF00300547)

96. Zirpoli JA, Black JM, Gabriel PO. 2013 Parasites and plumage in Steller's jays: an experimental field test of the parasite-mediated handicap hypothesis. *Ethol Ecol Evol* **25**, 103–116. (doi:10.1080/03949370.2012.742462)

97. Senar JC, Figuerola J, Pascual J. 2002 Brighter yellow blue tits make better parents. *Proc R Soc B* **269**, 257–261. (doi:10.1098/rspb.2001.1882)

98. Zuk M, Kim T, Robinson SI, Johnsen TS. 1998 Parasites influence social rank and morphology, but not mate choice, in female red junglefowl, *Gallus gallus*. *Anim Behav* **56**, 493–499. (doi:10.1006/anbe.1998.0807)

99. Grindstaff JL, Lovern MB, Burtka JL, Hallmark-Sharber A. 2012 Structural coloration signals condition, parental investment, and circulating hormone levels in Eastern bluebirds (*Sialia sialis*). *Journal of Comparative Physiology A* **198**, 625–637. (doi:10.1007/s00359-012-0735-0)

100. Setchel JM, Wickings EJ. 2004 Sexual swelling in mandrills (*Mandrillus sphinx*): a test of the reliable indicator hypothesis. *Behav Ecol* **15**, 438–445. (doi:10.1093/beheco/arh027)

101. Rigaill L, Higham JP, Lee PC, Blin A, Garcia C. 2013 Multimodal sexual signaling and mating behavior in olive baboons (*Papio anubis*). *Am J Primatol* **75**, 774–787. (doi:10.1002/ajp.22154)

102. Möhle U, Heistermann M, Dittami J, Reinberg V, Hodges JK. 2005 Patterns of anogenital swelling size and their endocrine correlates during ovulatory cycles and early pregnancy in free-ranging barbary macaques (*Macaca sylvanus*) of Gibraltar. *Am J Primatol* **66**, 351–368. (doi:10.1002/ajp.20161)

103. Huchard E, Courtiol A, Benavides JA, Knapp LA, Raymond M, Cowlishaw G. 2009 Can fertility signals lead to quality signals? Insights from the evolution of primate sexual swellings. *Proc R Soc B* **276**, 1889–1897. (doi:10.1098/rspb.2008.1923)

104. Brenowitz EA, Nalls B, Wingfield JC, Kroodsma DE. 1991 Seasonal changes in avian song nuclei without seasonal changes in song repertoire. *Journal of Neuroscience* **11**, 1367–1374. (doi:10.1523/jneurosci.11-05-01367.1991)

105. Airey DC, Buchanan KL, Szekely T, Catchpole CK, DeVoogd TJ. 2000 Song, sexual selection, and a song control nucleus (HVc) in the brains of European sedge warblers. *J Neurobiol* **44**, 1–6. (doi:10.1002/1097-4695(200007)44:1<1::AID-NEU1>3.0.CO;2-V)

106. Brenowitz EA, Lent K, Kroodsma DE. 1995 Brain space for learned song in birds develops independently of song learning. *Journal of Neuroscience* **15**, 6281–6286. (doi:10.1523/jneurosci.15-09-06281.1995)

107. Molnár O, Bajer K, Török J, Herczeg G. 2012 Individual quality and nuptial throat colour in male European green lizards. *J Zool* **287**, 233–239. (doi:10.1111/j.1469-7998.2012.00916.x)

108. Doucet SM, Montgomerie R. 2003 Multiple sexual ornaments in satin bowerbirds: ultraviolet plumage and bowers signal different aspects of male quality. *Behav Ecol* **14**, 503–509. (doi:10.1093/beheco/arg035)

109. Peters A, Kurvers RHJM, Roberts ML, Delhey K. 2011 No evidence for general condition-dependence of structural plumage colour in blue tits: an experiment. *J Evol Biol* **24**, 976–987. (doi:10.1111/j.1420-9101.2011.02229.x)

110. Roberts ML, Ras E, Peters A. 2009 Testosterone increases UV reflectance of sexually selected crown plumage in male blue tits. *Behav Ecol* **20**, 535–541. (doi:10.1093/beheco/arp028)

111. Peters A, Delhey K, Goymann W, Kempenaers B. 2006 Age-dependent association between testosterone and crown UV coloration in male blue tits (*Parus caeruleus*). *Behav Ecol Sociobiol* **59**, 666–673. (doi:10.1007/s00265-005-0095-7)

112. Vinkler M, Schnitzer J, Munclinger P, Albrecht T. 2012 Phytohaemagglutinin skin-swelling test in scarlet rosefinch males: low-quality birds respond more strongly. *Anim Behav* **83**, 17–23. (doi:10.1016/j.anbehav.2011.10.001)

113. Merilä J, Sheldon BC, Lindström K. 1999 Plumage brightness in relation to haematozoan infections in the greenfinch *Carduelis chloris*: bright males are a good bet. *Ecoscience* **6**, 12–18. (doi:10.1080/11956860.1999.11952203)

114. Martínez-Padilla J, Mougeot F, Pérez-Rodríguez L, Bortolotti GR. 2007 Nematode parasites reduce carotenoid-based signalling in male red grouse. *Biol Lett* **3**, 161–164. (doi:10.1098/rsbl.2006.0593)

115. Chui CKS, Mcgraw KJ, Doucet SM. 2011 Carotenoid-based plumage coloration in golden-crowned kinglets *Regulus satrapa*: pigment characterization and relationships with migratory timing and condition. *J Avian Biol* **42**, 309–322. (doi:10.1111/j.1600-048X.2011.05240.x)

116. Biard C, Saulnier N, Gaillard M, Moreau J. 2010 Carotenoid-based bill colour is an integrative signal of multiple parasite infection in blackbird. *Naturwissenschaften* **97**, 987–995. (doi:10.1007/s00114-010-0716-5)

117. Fenoglio S, Cucco M, Fracchia L, Martinotti MG, Malacarne G. 2004 Shield colours of the moorhen are differently related to bacterial presence and health parameters. *Ethol Ecol Evol* **16**, 171–180. (doi:10.1080/08927014.2004.9522645)

118. Rémy A, Grégoire A, Perret P, Doutrelant C. 2010 Mediating male-male interactions: the role of the UV blue crest coloration in blue tits. *Behav Ecol Sociobiol* **64**, 1839–1847. (doi:10.1007/s00265-010-0995-z)

119. Korsten P, Dijkstra TH, Komdeur J. 2007 Is UV signalling involved in male-male territorial conflict in the blue tit (*Cyanistes caeruleus*)? A new experimental approach. *Behaviour* **144**, 447–470. (doi:10.1163/156853907780756030)

120. Emery MA, Whitten PL. 2003 Size of sexual swellings reflects ovarian function in chimpanzees (*Pan troglodytes*). *Behav Ecol Sociobiol* **54**, 340–351. (doi:10.1007/s00265-003-0648-6)

121. Peters A, Delhey K, Johnsen A, Kempenaers B. 2007 The condition-dependent development of carotenoid-based and structural plumage in nestling blue tits: Males and females differ. *Am Nat* **169**. (doi:10.1086/510139)

122. Hadfield JD, Owens IPF. 2006 Strong environmental determination of a carotenoid-based plumage trait is not mediated by carotenoid availability. *J Evol Biol* **19**, 1104–1114. (doi:10.1111/j.1420-9101.2006.01095.x)

123. Dunn PO, Garvin JC, Whittingham LA, Freeman-Gallant CR, Hasselquist D. 2010 Carotenoid and melanin-based ornaments signal similar aspects of male quality in two populations of the common yellowthroat. *Funct Ecol* **24**, 149–158. (doi:10.1111/j.1365-2435.2009.01606.x)

124. Figuerola J, Muñoz E, Gutiérrez R, Ferrer D. 1999 Blood parasites, leucocytes and plumage brightness in the cirl bunting, *Emberiza cirlus*. *Funct Ecol* **13**, 594–601. (doi:10.1046/j.1365-2435.1999.00354.x)

125. Biard C, Surai PF, Møller AP. 2007 An analysis of pre- and post-hatching maternal effects mediated by carotenoids in the blue tit. *J Evol Biol* **20**, 326–339. (doi:10.1111/j.1420-9101.2006.01194.x)

126. Isaksson C, McLaughlin P, Monaghan P, Andersson S. 2007 Carotenoid pigmentation does not reflect total non-enzymatic antioxidant activity in plasma of adult and nestling great tits, *Parus major*. *Funct Ecol* **21**, 1123–1129. (doi:10.1111/j.1365-2435.2007.01317.x)

127. Eckert CG, Weatherhead PJ. 1987 Male characteristics, parental quality and the study of mate choice in the red-winged blackbird (*Agelaius phoeniceus*). *Behav Ecol Sociobiol* **20**, 35–42. (doi:10.1007/BF00292164)

128. Hõrak P, Sild E, Soomets U, Sepp T, Kilk K. 2010 Oxidative stress and information content of black and yellow plumage coloration: an experiment with greenfinches. *J Exp Biol* **213**, 2225–2233. (doi:10.1242/jeb.042085)

129. Mougeot F, Martínez-Padilla J, Blount JD, Pérez-Rodríguez L, Webster LMI, Piertney SB. 2010 Oxidative stress and the effect of parasites on a carotenoid-based ornament. *J Exp Biol* **213**, 400–407. (doi:10.1242/jeb.037101)

130. Pérez-Rodríguez L, Viñuela J. 2008 Carotenoid-based bill and eye ring coloration as honest signals of condition: an experimental test in the red-legged partridge (*Alectoris rufa*). *Naturwissenschaften* **95**, 821–830. (doi:10.1007/s00114-008-0389-5)

131. Keyser AJ, Hill GE. 2000 Structurally based plumage coloration is an honest signal of quality in male blue grosbeaks. *Behav Ecol* **11**, 202–209. (doi:10.1093/beheco/11.2.202)

132. Buchanan KL, Catchpole CK. 2000 Extra-pair paternity in the socially monogamous sedge warbler *Acrocephalus schoenobaenus* as revealed by multilocus DNA fingerprinting. *Ibis* **142**, 12–20. (doi:10.1111/j.1474-919x.2000.tb07678.x)

133. Veiga JP. 1993 Badge size, phenotypic quality, and reproductive success in the house sparrow: a study on honest advertisement. *Evolution* **47**, 1161–1170. (doi:10.1111/j.1558-5646.1993.tb02143.x)

134. Kose M, Mänd R, Møller AP. 1999 Sexual selection for white tail spots in the barn swallow in relation to habitat choice by feather lice. *Anim Behav* **58**, 1201–1205. (doi:10.1006/anbe.1999.1249)

135. Kose M, Møller AP. 1999 Sexual selection, feather breakage and parasites: the importance of white spots in the tail of the barn swallow (*Hirundo rustica*). *Behav Ecol Sociobiol* **45**, 430–436. (doi:10.1007/s002650050581)

136. Liker A, Barta Z. 2001 Male badge size predicts dominance against females in house sparrows. *Condor* **103**, 151–157. (doi:10.1093/condor/103.1.151)

137. Solberg EJ, Ringsby TH. 1997 Does male badge size signal status in small island populations of house sparrows, *Passer domesticus*? *Ethology* **103**, 177–186. (doi:10.1111/j.1439-0310.1997.tb00114.x)

138. Møller AP. 1987 Variation in badge size in male house sparrows *Passer domesticus*: evidence for status signalling. *Anim Behav* **35**, 1637–1644. (doi:10.1016/S0003-3472(87)80056-8)

139. Henderson LJ, Heidinger BJ, Evans NP, Arnold KE. 2013 Ultraviolet crown coloration in female blue tits predicts reproductive success and baseline corticosterone. *Behav Ecol* **24**, 1299–1305. (doi:10.1093/beheco/art066)

140. del Cerro S *et al.* 2010 Carotenoid-based plumage colouration is associated with blood parasite richness and stress protein levels in blue tits (*Cyanistes caeruleus*). *Oecologia* **162**, 825–835. (doi:10.1007/s00442-009-1510-y)

141. Garamszegi LZ, Hegyi G, Heylen D, Ninni P, De Lope F, Eens M, Møller AP. 2006 The design of complex sexual traits in male barn swallows: associations between signal attributes. *J Evol Biol* **19**, 2052–2066. (doi:10.1111/j.1420-9101.2006.01135.x)

142. Edler AU, Friedl TWP. 2010 Individual quality and carotenoid-based plumage ornaments in male red bishops (*Euplectes orix*): plumage is not all that counts. *Biol J Linn Soc* **99**, 384–397. (doi:10.1111/j.1095-8312.2009.01354.x)

143. Maney DL, Davis AK, Goode CT, Reid A, Showalter C. 2008 Carotenoid-based plumage coloration predicts leukocyte parameters during the breeding season in northern cardinals (*Cardinalis cardinalis*). *Ethology* **114**, 369–380. (doi:10.1111/j.1439-0310.2008.01476.x)

144. Dufva R, Allander K. 1995 Intraspecific variation in plumage coloration reflects immune response in great tit (*Parus major*) males. *Funct Ecol* **9**, 785. (doi:10.2307/2390253)

145. Wiehn J, Korpimáki E, Bildstein KL, Sorjonen J. 1997 Mate choice and reproductive success in the American kestrel: a role for blood parasites? *Ethology* **103**, 304–317. (doi:10.1111/j.1439-0310.1997.tb00020.x)

146. Thompson CW, Hillgarth N, Leu M, McClure HE. 1997 High parasite load in house finches (*Carpodacus mexicanus*) is correlated with reduced expression of a sexually selected trait. *Am Nat* **149**, 270–294. (doi:10.1086/285990)

147. Seutin G. 1994 Plumage redness in redpoll finches does not reflect hemoparasitic infection. *Oikos* **70**, 280. (doi:10.2307/3545639)

148. Weatherhead PJ, Metz KJ, Bennett GF, Irwin RE. 1993 Parasite faunas, testosterone and secondary sexual traits in male red-winged blackbirds. *Behav Ecol Sociobiol* **33**, 13–23. (doi:10.1007/BF00164342)

149. Vergara P, Fargallo JA, MartÍnez-Padilla J, Lemus JA. 2009 Inter-annual variation and information content of melanin-based coloration in female Eurasian kestrels. *Biol J Linn Soc* **97**, 781–790. (doi:10.1111/j.1095-8312.2009.01263.x)

150. Tella JL, Forero MG, Donázar JA, Hiraldo F. 1997 Is the expressin of male traits in female lesser kestrels related to sexual selection. *Ethology* **103**, 72–81. (doi:10.1111/j.1439-0310.1997.tb00008.x)

151. Møller AP. 1993 Female preference for apparently symmetrical male sexual ornaments in the barn swallow *Hirundo rustica*. *Behav Ecol Sociobiol* **32**, 371–376. (doi:10.1007/BF00168820)

152. Brommer JE, Ahola K, Karstinen T. 2005 The colour of fitness: plumage coloration and lifetime reproductive success in the tawny owl. *Proc R Soc B* **272**, 935–940. (doi:10.1098/rspb.2005.3052)

153. Horne TJ. 1998 Evolution of females choice in the bank vole. University of Jyväskylä.

154. De Lope F, Møller AP. 1993 Female reproductive effort depends on the degree of ornamentation of their mates. *Evolution* **47**, 1152–1160. (doi:10.1111/j.1558-5646.1993.tb02142.x)

155. Møller AP. 1993 Sexual selection in the barn swallow *Hirundo rustica*. III. Female tail ornaments. *Evolution* **47**, 417. (doi:10.2307/2410061)

156. Welch AM, Semlitsch RD, Gerhardt HC. 1998 Call duration as an indicator of genetic quality in male gray tree frogs. *Science* **280**, 1928–1930. (doi:10.1126/science.280.5371.1928)

157. Alatalo R V., Kotiaho J, Mappes J, Parri S. 1998 Mate choice for offspring performance: major benefits or minor costs? *Proc R Soc B* **265**, 2297–2301. (doi:10.1098/rspb.1998.0574)

158. Hoikkala A, Suvanto L. 1999 Male courtship song frequency as an indicator of male mating success in *Drosophila montana*. *J Insect Behav* **12**, 599–609. (doi:10.1023/A:1020971518710)

159. Hasselquist D, Bensch S, von Schantz T. 1996 Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature* **381**, 229–232. (doi:10.1038/381229a0)

160. Freeman-Gallant CR, Schneider RL, Taff CC, Dunn PO, Whittingham LA. 2014 Contrasting patterns of selection on the size and coloration of a female plumage ornament in common yellowthroats. *J Evol Biol* **27**, 982–991. (doi:10.1111/jeb.12369)

161. Gladbach A, Gladbach DJ, Kempenaers B, Quillfeldt P. 2010 Female-specific colouration, carotenoids and reproductive investment in a dichromatic species, the upland goose *Chloephaga picta leucoptera*. *Behav Ecol Sociobiol* **64**, 1779–1789. (doi:10.1007/s00265-010-0990-4)

162. Quinard A, Cézilly F, Motreuil S, Rossi JM, Biard C. 2017 Reduced sexual dichromatism, mutual ornamentation, and individual quality in the monogamous Zenaida dove *Zenaida aurita*. *J Avian Biol* **48**, 489–501. (doi:10.1111/jav.00902)

163. Simons MJP, Briga M, Koetsier E, Folkertsma R, Wubs MD, Dijkstra C, Verhulst S. 2012 Bill redness is positively associated with reproduction and survival in male and female zebra finches. *PLoS One* **7**. (doi:10.1371/journal.pone.0040721)

164. Komdeur J, Oorebeek M, Van Overveld T, Cuthill IC. 2005 Mutual ornamentation, age, and reproductive performance in the European starling. *Behav Ecol* **16**, 805–817. (doi:10.1093/beheco/ari059)

165. Daunt F, Monaghan P, Wanless S, Harris MP. 2003 Sexual ornament size and breeding performance in female and male European shags *Phalacrocorax aristotelis*. *Ibis* **145**, 54–60. (doi:10.1046/j.1474-919X.2003.00120.x)

166. Morrison A, Flood NJ, Reudink MW. 2014 Reproductive correlates of plumage coloration of female mountain bluebirds. *J Field Ornithol* **85**, 168–179. (doi:10.1111/jofo.12058)

167. Mänd R, Tilgar V, Møller AP. 2005 Negative relationship between plumage colour and breeding output in female great tits, *Parus major*. *Evol Ecol Res* **7**, 1013–1023.

168. Muma KE, Weatherhead PJ. 1989 Male traits expressed in females: direct or indirect sexual selection? *Behav Ecol Sociobiol* **25**, 23–31. (doi:10.1007/BF00299707)

169. Beck ML, Hopkins WA, Hawley DM. 2015 Relationships among plumage coloration, blood selenium concentrations and immune responses of adult and nestling tree swallows. *J Exp Biol* **218**, 3415–3424. (doi:10.1242/jeb.123794)

170. Potti J, Merino S. 1996 Decreased levels of blood trypanosome infection correlate with female expression of a male secondary sexual trait: implications for sexual selection. *Proc R Soc B* **263**, 1199–1204. (doi:10.1098/rspb.1996.0176)

171. Bentz AB, Siefferman L. 2013 Age-dependent relationships between coloration and reproduction in a species exhibiting delayed plumage maturation in females. *J Avian Biol* **44**, 080–088. (doi:10.1111/j.1600-048X.2012.05730.x)

172. Osmond MM, Reudink MW, Germain RR, Marra PP, Nocera JJ, Boag PT, Ratcliffe LM. 2013 Relationships between carotenoid-based female plumage and age, reproduction, and mate colour in the American redstart (*Setophaga ruticilla*). *Can J Zool* **91**, 589–595. (doi:10.1139/cjz-2013-0017)

173. Bulluck LP, Foster MJ, Kay S, Cox DE, Viverette C, Huber S. 2017 Feather carotenoid content is correlated with reproductive success and provisioning rate in female prothonotary warblers. *Auk* **134**, 229–239. (doi:10.1642/AUK-16-151.1)

174. Doutrelant C, Grégoire A, Grnac N, Gomez D, Lambrechts MM, Perret P. 2008 Female coloration indicates female reproductive capacity in blue tits. *J Evol Biol* **21**, 226–233. (doi:10.1111/j.1420-9101.2007.01451.x)

175. Cuervo JJ, Møller AP, De Lope F. 2003 Experimental manipulation of tail length in female barn swallows (*Hirundo rustica*) affects their future reproductive success. *Behav Ecol* **14**, 451–456. (doi:10.1093/beheco/arg027)

176. Silva N, Avilés JM, Danchin E, Parejo D. 2008 Informative content of multiple plumage-coloured traits in female and male European rollers. *Behav Ecol Sociobiol* **62**, 1969–1979. (doi:10.1007/s00265-008-0628-y)

177. Viblanc VA *et al.* 2016 Mutually honest? Physiological 'qualities' signalled by colour ornaments in monomorphic king penguins. *Biol J Linn Soc* **118**, 200–214. (doi:10.1111/bij.12729)

178. van Dijk RE, Robles R, Groothuis TGG, de Vries B, Eising CM. 2015 Reproductive effort of both male and female bar-throated apalis *Apalis thoracica* is predicted by ornamentation of self and mate. *Ibis* **157**, 731–742. (doi:10.1111/ibi.12279)

179. Lindström KM, Hasselquist D, Wikelski M. 2005 House sparrows (*Passer domesticus*) adjust their social status position to their physiological costs. *Horm Behav* **48**, 311–320. (doi:10.1016/j.yhbeh.2005.04.002)

180. Hein WK, Westneat DF, Poston JP. 2003 Sex of opponent influences response to a potential status signal in house sparrows. *Anim Behav* **65**, 1211–1221. (doi:10.1006/anbe.2003.2132)

181. Riters L V., Teague DP, Schroeder MB. 2004 Social status interacts with badge size and neuroendocrine physiology to influence sexual behavior in male house sparrows (*Passer domesticus*). *Brain Behav Evol* **63**, 141–150. (doi:10.1159/000076240)

182. Gonzalez G, Sorci G, Smith LC, De Lope F. 2002 Social control and physiological cost of cheating in status signalling male house sparrows (*Passer domesticus*). *Ethology* **108**, 289–302. (doi:10.1046/j.1439-0310.2002.00779.x)

183. Bókony V, Lendvai ÁZ, Liker A. 2006 Multiple cues in status signalling: the role of wingbars in aggressive interactions of male house sparrows. *Ethology* **112**, 947–954. (doi:10.1111/j.1439-0310.2006.01246.x)

184. Delhey K, Kempenaers B. 2006 Age differences in blue tit *Parus caeruleus* plumage colour: within-individual changes or colour-biased survival? *J Avian Biol* **37**, 339–348. (doi:10.1111/j.2006.0908-8857.03655.x)

185. Surmacki A, Stępniewski J, Stępniewska M. 2015 Juvenile sexual dimorphism, dichromatism and condition-dependent signaling in a bird species with early pair bonds. *J Ornithol* **156**, 65–73. (doi:10.1007/s10336-014-1108-y)

186. Hidalgo-Garcia S. 2006 The carotenoid-based plumage coloration of adult blue tits *Cyanistes caeruleus* correlates with the health status of their brood. *Ibis* **148**, 727–734. (doi:10.1111/j.1474-919X.2006.00575.x)

187. Mougeot F, Redpath SM, Leckie F. 2005 Ultra-violet reflectance of male and female red grouse, *Lagopus lagopus scoticus*: sexual ornaments reflect nematode parasite intensity. *J Avian Biol* **36**, 203–209. (doi:10.1111/j.0908-8857.2005.03424.x)

188. Saino N *et al.* 2013 Viability is associated with melanin-based coloration in the barn swallow (*Hirundo rustica*). *PLoS One* **8**, e60426. (doi:10.1371/journal.pone.0060426)

189. Vortman Y, Lotem A, Dor R, Lovette IJ, Safran RJ. 2011 The sexual signals of the East-Mediterranean barn swallow: a different swallow tale. *Beh Ecol* **22**, 1344–1352. (doi:10.1093/beheco/arr139)

190. Faivre B, Grégoire A, Préault M, Cézilly F, Sorci G. 2003 Immune activation rapidly mirrored in a secondary sexual trait. *Science* **300**, 103–103. (doi:10.1126/science.1081802)

191. Grunst AS, Rotenberry JT, Grunst ML. 2014 Age-dependent relationships between multiple sexual pigments and condition in males and females. *Behav Ecol* **25**, 276–287. (doi:10.1093/beheco/art124)

192. Kraaijeveld K, Gregurke J, Hall C, Komdeur J, Mulder RA. 2004 Mutual ornamentation, sexual selection, and social dominance in the black swan. *Behav Ecol* **15**, 380–389. (doi:10.1093/beheco/arh023)

193. Tschirren B, Fitze PS, Richner H. 2003 Proximate mechanisms of variation in the carotenoid-based plumage coloration of nestling great tits (*Parus major* L.). *J Evol Biol* **16**, 91–100. (doi:10.1046/j.1420-9101.2003.00483.x)

194. Figuerola J, Domènech J, Senar JC. 2003 Plumage colour is related to ectosymbiont load during moult in the serin, *Serinus serinus*: an experimental study. *Anim Behav* **65**, 551–557. (doi:10.1006/anbe.2003.2072)

195. De Kogel CH, Prijs HJ. 1996 Effects of brood size manipulations on sexual attractiveness of offspring in the zebra finch. *Anim Behav* **51**, 699–708. (doi:10.1006/anbe.1996.0073)

196. McGraw KJ, Hill GE. 2000 Differential effects of endoparasitism on the expression of carotenoid- and melanin-based ornamental coloration. *P Proc R Soc B* **267**, 1525–1531. (doi:10.1098/rspb.2000.1174)

197. Griffith SC, Owens IPF, Burke T. 1999 Female choice and annual reproductive success favour less-ornamented male house sparrows. *Proc R Soc B* **266**, 765–770. (doi:10.1098/rspb.1999.0703)

198. McGraw KJ, Mackillop EA, Dale J, Hauber ME. 2002 Different colors reveal different information: how nutritional stress affects the expression of melanin- and structurally based ornamental plumage. *J Exp Biol* **205**, 3747–3755. (doi:10.1242/jeb.205.23.3747)

199. Fitze PS, Richner H. 2002 Differential effects of a parasite on ornamental structures based on melanins and carotenoids. *Behav Ecol* **13**, 401–407. (doi:10.1093/beheco/13.3.401)

200. Johnsen A, Delhey K, Schlicht E, Peters A, Kempenaers B. 2005 Male sexual attractiveness and parental effort in blue tits: a test of the differential allocation hypothesis. *Anim Behav* **70**, 877–888. (doi:10.1016/j.anbehav.2005.01.005)

201. Maguire SE, Safran RJ. 2010 Morphological and genetic predictors of parental care in the North American barn swallow *Hirundo rustica erythrogaster*. *J Avian Biol* **41**, 74–82. (doi:10.1111/j.1600-048X.2009.04806.x)

202. Sundberg J, Larsson C. 1994 Male coloration as an indicator of parental quality in the yellowhammer, *Emberiza citrinella*. *Anim Behav*. **48**, 885–892. (doi:10.1006/anbe.1994.1313)

203. Lozano GA, Lemon RE. 1996 Male plumage, paternal care and reproductive success in yellow warblers, *Dendroica petechia*. *Anim Behav* **51**, 265–272. (doi:10.1006/anbe.1996.0027)

204. Studd M V., Robertson RJ. 1985 Sexual selection and variation in reproductive strategy in male yellow warblers (*Dendroica petechia*). *Behav Ecol Sociobiol* **17**, 101–109. (doi:10.1007/BF00299242)

205. Møller AP. 1994 Sexual selection in the barn swallow (*Hirundo rustica*). IV. Patterns of fluctuating asymmetry and selection against asymmetry. *Evolution* **48**, 658. (doi:10.2307/2410476)

206. Jawor JM, Gray N, Beall SM, Breitwisch R. 2004 Multiple ornaments correlate with aspects of condition and behaviour in female northern cardinals, *Cardinalis cardinalis*. *Anim Behav* **67**, 875–882. (doi:10.1016/j.anbehav.2003.05.015)

207. Linville SU, Breitwisch R, Schilling AJ. 1998 Plumage brightness as an indicator of parental care in northern cardinals. *Anim Behav* **55**, 119–127. (doi:10.1006/anbe.1997.0595)

208. Hõrak P, Saks L, Karu U, Ots I, Surai PF, McGraw KJ. 2004 How coccidian parasites affect health and appearance of greenfinches. *Journal of Animal Ecology* **73**, 935–947. (doi:10.1111/j.0021-8790.2004.00870.x)

209. Møller AP, Tegelström H. 1997 Extra-pair paternity and tail ornamentation in the barn swallow *Hirundo rustica*. *Behav Ecol Sociobiol* **41**, 353–360. (doi:10.1007/s002650050395)

210. Møller AP. 1994 Male ornament size as a reliable cue to enhanced offspring viability in the barn swallow. *Proc Natl Acad Sci U S A* **91**, 6929–6932. (doi:10.1073/pnas.91.15.6929)

211. Møller AP. 1994 Phenotype-dependent arrival time and its consequences in a migratory bird. *Behav Ecol Sociobiol* **35**, 115–122. (doi:10.1007/BF00171501)

212. Møller AP. 1994 Symmetrical male sexual ornaments, paternal care, and offspring quality. *Behav Ecol* **5**, 188–194. (doi:10.1093/beheco/5.2.188)

213. Burley N, Coopersmith CB. 1987 Bill color preferences of zebra finches. *Ethology* **76**, 133–151. (doi:10.1111/j.1439-0310.1987.tb00679.x)

214. Møller AP. 1992 Female swallow preference for symmetrical male sexual ornaments. *Nature* **357**, 238–240. (doi:10.1038/357238a0)

215. Møller AP. 1992 Sexual selection in the monogamous barn swallow (*Hirundo rustica*). II. Mechanisms of sexual selection. *J Evol Biol* **5**, 603–624. (doi:10.1046/j.1420-9101.1992.5040603.x)

216. Møller AP. 1990 Male tail length and female mate choice in the monogamous swallow *Hirundo rustica*. *Anim Behav* **39**, 458–465. (doi:10.1016/S0003-3472(05)80409-9)

217. Rintamäki PT, Alatalo R V., Höglund J, Lundberg A. 1997 Fluctuating asymmetry and copulation success in lekking black grouse. *Anim Behav* **54**, 265–269. (doi:10.1006/anbe.1996.0434)

218. Collins SA, Hubbard C, Houtman AM. 1994 Female mate choice in the zebra finch - the effect of male beak colour and male song. *Behav Ecol Sociobiol* **35**, 21–25. (doi:10.1007/BF00167055)

219. Møller AP. 1989 Viability costs of male tail ornaments in a swallow. *Nature* **339**, 132–135. (doi:10.1038/339132a0)

220. Møller AP. 1989 Natural and sexual selection on a plumage signal of status and on morphology in house sparrows, *Passer domesticus*. *J Evol Biol* **2**, 125–140. (doi:10.1046/j.1420-9101.1989.2020125.x)

221. Norris KJ. 1990 Female choice and the evolution of the conspicuous plumage coloration of monogamous male great tits. *Behav Ecol Sociobiol* **26**, 129–138. (doi:10.1007/BF00171582)

222. Evans MatthewR, Hatchwell BJ. 1992 An experimental study of male adornment in the scarlet-tufted malachite sunbird: I. The role of pectoral tufts in territorial defence. *Behav Ecol Sociobiol* **29**, 421–427. (doi:10.1007/BF00170171)

223. Saino N, Bolzern AM, Møller AP. 1997 Immunocompetence, ornamentation, and viability of male barn swallows (*Hirundo rustica*). *Proc Natl Acad Sci U S A* **94**, 549–552. (doi:10.1073/pnas.94.2.549)

224. Baeta R, Faivre B, Motreuil S, Gaillard M, Moreau J. 2008 Carotenoid trade-off between parasitic resistance and sexual display: an experimental study in the blackbird (*Turdus merula*). *Proc R Soc B* **275**, 427–434. (doi:10.1098/rspb.2007.1383)

225. Stirnemann I, Johnston G, Rich B, Robertson J, Kleindorfer S. 2009 Phytohaemagglutinin (PHA) response and bill-hue wavelength increase with carotenoid supplementation in diamond firetails (*Stagonopleura guttata*). *Emu* **109**, 344–351. (doi:10.1071/MU08069)

226. Hill GE, Hood WR, Huggins K. 2009 A multifactorial test of the effects of carotenoid access, food intake and parasite load on the production of ornamental feathers and bill coloration in American goldfinches. *J Exp Biol* **212**, 1225–1233. (doi:10.1242/jeb.026963)

227. Navara KJ, Hill GE. 2003 Dietary carotenoid pigments and immune function in a songbird with extensive carotenoid-based plumage coloration. *Behav Ecol* **14**, 909–916. (doi:10.1093/beheco/arg085)

228. Peters A, Delhey K, Andersson S, Van Noordwijk H, Förschler MI. 2008 Condition-dependence of multiple carotenoid-based plumage traits: an experimental study. *Funct Ecol* **22**, 831–839. (doi:10.1111/j.1365-2435.2008.01437.x)

229. Karu U, Saks L, Hõrak P. 2007 Carotenoid coloration in greenfinches is individually consistent irrespective of foraging ability. *Physiol Biochem Zool* **80**, 663–670. (doi:10.1086/521084)

230. Blount JD, Metcalfe NB, Birkhead TR, Surai PF. 2003 Carotenoid modulation of immune function and sexual attractiveness in zebra finches. *Science* **300**, 125–127. (doi:10.1126/science.1082142)

231. Ryan MJ, Perrill SA, Wilczynski W. 1992 Auditory tuning and call frequency predict population-based mating preferences in the cricket frog, *Acris crepitans*. *Am Nat* **139**, 1370–1383. (doi:10.1086/285391)

232. McGraw KJ, Ardia DR. 2003 Carotenoids, immunocompetence, and the information content of sexual colors: an experimental test. *Am Nat* **162**, 704–712. (doi:10.1086/378904)

233. Jensen H, Sæther BE, Ringsby TH, Tufto J, Griffith SC, Ellegren H. 2004 Lifetime reproductive success in relation to morphology in the house sparrow *Passer domesticus*. *Journal of Animal Ecology* **73**, 599–611. (doi:10.1111/j.0021-8790.2004.00837.x)

234. Václav R, Hoi H. 2002 Different reproductive tactics in house sparrows signalled by badge size: is there a benefit to being average? *Ethology* **108**, 569–582. (doi:10.1046/j.1439-0310.2002.00799.x)

235. Birkhead ATR, Fletcher F. 1995 Male phenotype and ejaculate quality in the zebra finch *Taeniopygia guttata*. *Proc R Soc B* **262**, 329–334. (doi:10.1098/rspb.1995.0213)

236. Potti J, Montalvo S. 1991 Male arrival and female mate choice in pied flycatchers *Ficedula hypoleuca* in central Spain. *Ornis Scandinavica* **22**, 45–54. (doi:10.2307/3676620)

237. Evans JP. 2010 Quantitative genetic evidence that males trade attractiveness for ejaculate quality in guppies. *Proc R Soc B* **277**, 3195–3201. (doi:10.1098/rspb.2010.0826)

238. Engen F, Folstad I. 1999 Cod courtship song: a song at the expense of dance? *Can J Zool* **77**, 542–550. (doi:10.1139/cjz-77-4-542)

239. Peters A, Denk AG, Delhey K, Kempenaers B. 2004 Carotenoid-based bill colour as an indicator of immunocompetence and sperm performance in male mallards. *J Evol Biol* **17**, 1111–1120. (doi:10.1111/j.1420-9101.2004.00743.x)

240. Hasselquist D. 1998 Polygyny in great reed warblers: a long-term study of factors contributing to male fitness. *Ecology* **79**, 2376–2390. (doi:10.1890/0012-9658(1998)079[2376:PIGRWA]2.0.CO;2)

241. Roulin A, Dijkstra C, Riols C, Ducrest AL. 2001 Female- and male-specific signals of quality in the barn owl. *J Evol Biol* **14**, 255–266. (doi:10.1046/j.1420-9101.2001.00274.x)

242. Roulin A, Riols C, Dijkstra C, Ducrest AL. 2001 Female plumage spottiness signals parasite resistance in the barn owl (*Tyto alba*). *Behav Ecol* **12**, 103–110. (doi:10.1093/oxfordjournals.beheco.a000371)

243. Siefferman L, Hill GE. 2003 Structural and melanin coloration indicate parental effort and reproductive success in male eastern bluebirds. *Behav Ecol* **14**, 855–861. (doi:10.1093/beheco/arg063)

244. Siefferman L, Hill GE. 2005 Evidence for sexual selection on structural plumage coloration in female eastern bluebirds (*Sialia sialis*). *Evolution* **59**, 1819–1828. (doi:10.1111/j.0014-3820.2005.tb01828.x)

245. Balenger SL, Scott Johnson L, Brubaker JL, Ostlind E. 2007 Parental effort in relation to structural plumage coloration in the mountain bluebird (*Sialia currucoides*). *Ethology* **113**, 838–846. (doi:10.1111/j.1439-0310.2007.01386.x)

246. McGregor PK, Krebs JR, Perrins CM. 1981 Song repertoires and lifetime reproductive success in the great tit (*Parus major*). *Am Nat* **118**, 149–159. (doi:10.1086/283811)

247. Voltura KM, Schwagmeyer PL, Mock DW. 2002 Parental feeding rates in the house sparrow, *Passer domesticus*: are larger-badged males better fathers? *Ethology* **108**, 1011–1022. (doi:10.1046/j.1439-0310.2002.00831.x)

248. Nakagawa S, Ockendon N, Gillespie DOS, Hatchwell BJ, Burke T. 2007 Does the badge of status influence parental care and investment in house sparrows? An experimental test. *Oecologia* **153**, 749–760. (doi:10.1007/s00442-007-0765-4)

249. Møller AP. 1991 Parasite load reduces song output in a passerine bird. *Anim Behav* **41**, 723–730. (doi:10.1016/S0003-3472(05)80909-1)

250. Alonso-Alvarez C, Bertrand S, Devevey G, Gaillard M, Prost J, Faivre B, Sorci G. 2004 An experimental test of the dose-dependent effect of carotenoids and immune activation on sexual signals and antioxidant activity. *Am Nat* **164**, 651–659. (doi:10.1086/424971)

251. Lambrechts M, Dhondt AA. 1986 Male quality, reproduction, and survival in the great tit (*Parus major*). *Behav Ecol Sociobiol* **19**, 57–63. (doi:10.1007/BF00303843)

252. Møller AP. 1991 Viability is positively related to degree of ornamentation in male swallows. *Proc R Soc B* **243**, 145–148. (doi:10.1098/rspb.1991.0023)

253. Møller AP, Nielsen JT. 1997 Differential predation cost of a secondary sexual character: sparrowhawk predation on barn swallows. *Anim Behav* **54**, 1545–1551. (doi:10.1006/anbe.1997.9998)

254. Brown CR, Brown MB. 1999 Natural selection on tail and bill morphology in barn swallows *Hirundo rustica* during severe weather. *Ibis* **141**, 652–659. (doi:10.1111/j.1474-919x.1999.tb07373.x)

255. Møller AP, de Lope F. 1994 Differential costs of a secondary sexual character: an experimental test of the handicap principle. *Evolution* **48**, 1676. (doi:10.2307/2410256)

256. Buchanan KL, Catchpole CK. 2000 Song as an indicator of male parental effort in the sedge warbler. *Proc R Soc B* **267**, 321–326. (doi:10.1098/rspb.2000.1003)

257. Birkhead TR, Fletcher F, Pellatt EJ. 1998 Sexual selection in the zebra finch *Taeniopygia guttata*: condition, sex traits and immune capacity. *Behav Ecol Sociobiol* **44**, 179–191. (doi:10.1007/s002650050530)

258. Adamo SA, Kovalko I, Easy RH, Stoltz D. 2014 A viral aphrodisiac in the cricket *Gryllus texensis*. *J Exp Biol* **217**, 1970–1976. (doi:10.1242/jeb.103408)

259. Smith MJ, Roberts JD. 2003 Call structure may affect male mating success in the quacking frog *Crinia georgiana* (Anura: Myobatrachidae). *Behav Ecol Sociobiol* **53**, 221–226. (doi:10.1007/s00265-002-0563-2)

260. Kingma SA, Szentirmai I, Székely T, Bókony V, Bleeker M, Liker A, Komdeur J. 2008 Sexual selection and the function of a melanin-based plumage ornament in polygamous penduline tits *Remiz pendulinus*. *Behav Ecol Sociobiol* **62**, 1277–1288. (doi:10.1007/s00265-008-0556-x)

261. David Ligon J, Zwartjes PW. 1995 Female red junglefowl choose to mate with multiple males. *Anim Behav* **49**, 127–135. (doi:10.1016/0003-3472(95)80160-X)

262. Ligon JD, Zwartjes PW. 1995 Ornate plumage of male red junglefowl does not influence mate choice by females. *Anim Behav* **49**, 117–125. (doi:10.1016/0003-3472(95)80159-6)

263. Kempenaers B, Verheyen GR, Dhondt AA. 1997 Extrapair paternity in the blue tit (*Parus caeruleus*): female choice, male characteristics, and offspring quality. *Behav Ecol* **8**, 481–492. (doi:10.1093/beheco/8.5.481)

264. Leisler B. 2000 Variation in extra-pair paternity in the polygynous great reed warbler (*Acrocephalus arundinaceus*). *J für Ornithol* **141**, 77. (doi:10.1046/j.1439-0361.2000.00047.x)

265. Chappell MA, Zuk M, Johnsen TS, Kwan TH. 1997 Mate choice and aerobic capacity in red junglefowl. *Behaviour* **134**, 511–529. (doi:10.1163/156853997X00502)

266. Kennedy CEJ, Endler JA, Poynton SL, McMinn H. 1987 Parasite load predicts mate choice in guppies. *Behav Ecol Sociobiol* **21**, 291–295. (doi:10.1007/BF00299966)

267. Hall ZJ, Bertin M, Bailey IE, Meddle SL, Healy SD. 2014 Neural correlates of nesting behavior in zebra finches (*Taeniopygia guttata*). *Behavioural Brain Research* **264**, 26–33. (doi:10.1016/j.bbr.2014.01.043)

268. Bernard DJ, Eens M, Ball GF. 1996 Age- and behavior-related variation in volumes of song control nuclei in male European starlings. *J Neurobiol* **30**, 329–339. (doi:10.1002/(SICI)1097-4695(199607)30:3<329::AID-NEU2>3.0.CO;2-6)

269. Taylor MI, Turner GF, Robinson RL, Stauffer JR. 1998 Sexual selection, parasites and bower height skew in a bower-building cichlid fish. *Anim Behav* **56**, 379–384. (doi:10.1006/anbe.1998.0795)

270. Pruett-jones SG, Pruett-jones MA, Jones HI. 1990 Parasites and sexual selection in birds of paradise. *Integr Comp Biol* **30**, 287–298. (doi:10.1093/icb/30.2.287)

271. Price DK, Burley NT. 1994 Constraints on the evolution of attractive traits: selection in male and female zebra finches. *Am Nat* **144**, 908–934. (doi:10.1086/285718)

272. Sundberg J. 1995 Parasites, plumage coloration and reproductive success in the yellowhammer, *Emberiza citrinella*. *Oikos* **74**, 331. (doi:10.2307/3545664)

273. Bosholn M, Fecchio A, Silveira P, Braga ÉM, Anciães M. 2016 Effects of avian malaria on male behaviour and female visitation in lekking blue-crowned manakins. *J Avian Biol* **47**, 457–465. (doi:10.1111/jav.00864)

274. Arak A. 1988 Female mate selection in the natterjack toad: active choice or passive atraction? *Behav Ecol Sociobiol* **22**, 317–327. (doi:10.1007/BF00295100)

275. Palokangas P, Korpimäki E, Hakkarainen H, Huhta E, Tolonen P, Alatalo R V. 1994 Female kestrels gain reproductive success by choosing brightly ornamented males. *Anim Behav* **47**, 443–448. (doi:10.1006/anbe.1994.1058)

276. Wiehn J. 1997 Plumage characteristics as an indicator of male parental quality in the American kestrel. *J Avian Biol* **28**, 47. (doi:10.2307/3677093)

277. Weatherhead PJ, Boag PT. 1995 Pair and extra-pair mating success relative to male quality in red-winged blackbirds. *Behav Ecol Sociobiol* **37**, 81–91. (doi:10.1007/BF00164153)

278. Hill GE. 1990 Female house finches prefer colourful males: sexual selection for a condition-dependent trait. *Anim Behav* **40**, 563–572. (doi:10.1016/S0003-3472(05)80537-8)

279. Andersson S. 1992 Female preference for long tails in lekking Jackson's widowbirds: experimental evidence. *Anim Behav* **43**, 379–388. (doi:10.1016/S0003-3472(05)80098-3)

280. Møller AP. 1988 Female choice selects for male sexual tail ornaments in the monogamous swallow. *Nature* **332**, 640–642. (doi:10.1038/332640a0)

281. Grant BR, Grant PR. 1987 Mate choice in Darwin's Finches. *Biol J Linn Soc* **32**, 247–270. (doi:10.1111/j.1095-8312.1987.tb00432.x)

282. Petrie M, Halliday T. 1994 Experimental and natural changes in the peacock's (*Pavo cristatus*) train can affect mating success. *Behav Ecol Sociobiol* **35**, 213–217. (doi:10.1007/BF00167962)

283. Andersson S. 1989 Sexual selection and cues for female choice in leks of Jackson's widowbird *Euplectes jacksoni*. *Behav Ecol Sociobiol* **25**, 403–410. (doi:10.1007/BF00300186)

284. Smith HG, Montgomerie R. 1991 Sexual selection and the tail ornaments of North American barn swallows. *Behav Ecol Sociobiol* **28**, 195–201. (doi:10.1007/BF00172171)

285. Grant BR. 1990 The significance of subadult plumage in Darwin's finches, *Geospiza fortis*. *Behav Ecol* **1**, 161–170. (doi:10.1093/beheco/1.2.161)

286. Røskaft E, Järvi T. 1983 Male plumage colour and mate choice of female pied flycatchers *Ficedula hypoleuca*. *Ibis* **125**, 396–400. (doi:10.1111/j.1474-919X.1983.tb03129.x)

287. Zuk M, Ligon JD, Thornhill R. 1992 Effects of experimental manipulation of male secondary sex characters on female mate preference in red jungle fowl. *Anim Behav* **44**, 999–1006. (doi:10.1016/S0003-3472(05)80312-4)

288. Gustafsson L, Qvarnström A, Sheldon BC. 1995 Trade-offs between life-history traits and a secondary sexual character in male collared flycatchers. *Nature* **375**, 311–313. (doi:10.1038/375311a0)

289. Järvi T, Røskaft E, Bakken M, Zumsteg B. 1987 Evolution of variation in male secondary sexual characteristics. *Behav Ecol Sociobiol* **20**, 161–169. (doi:10.1007/BF00299729)

290. McLennan DA, Shires VL. 1995 Correlation between the Level of infection with *Bunodera inconstans* and *Neoechinorhynchus rutili* and behavioral intensity in female brook sticklebacks. *J Parasitol* **81**, 675. (doi:10.2307/3283953)

291. Tárano Z. 2001 Variation in male advertisement calls in the Neotropical frog *Physalaemus enesefae*. *Copeia* **2001**, 1064–1072. (doi:10.1643/0045-8511(2001)001[1064:VIMACI]2.0.CO;2)

292. Isaksson C, Ornborg J, Prager M, Andersson S. 2008 Sex and age differences in reflectance and biochemistry of carotenoid-based colour variation in the great tit *Parus major*. *Biol J Linn Soc* **95**, 758–765. (doi:10.1111/j.1095-8312.2008.01033.x)

293. Zuk M, Thornhill R, Ligon JD, Johnson K, Austad S, Ligon SH, Thornhill NW, Costin C. 1990 The role of male ornaments and courtship behavior in female mate choice of red jungle fowl. *Am Nat* **136**, 459–473. (doi:10.1086/285107)

294. Houde AE, Torio AJ. 1992 Effect of parasitic infection on male color pattern and female choice in guppies. *Behav Ecol* **3**, 346–351. (doi:10.1093/beheco/3.4.346)

295. Greenspan SE, Roznik EA, Schwarzkopf L, Alford RA, Pike DA. 2016 Robust calling performance in frogs infected by a deadly fungal pathogen. *Ecol Evol* **6**, 5964–5972. (doi:10.1002/ece3.2256)

296. Gibson RM. 1990 Relationships between blood parasites, mating success and phenotypic cues in male sage grouse *Centrocercus urophasianus*. *Am Zool* **30**, 271–278. (doi:10.1093/icb/30.2.271)

297. Hill GE. 1991 Plumage coloration is a sexually selected indicator of male quality. *Nature* **350**, 337–339. (doi:10.1038/350337a0)

298. Fiske P, Kàlàs JA, Saether SA. 1994 Correlates of male mating success in the lekking great snipe (*Gallinago media*): results from a four-year study. *Behav Ecol* **5**, 210–218. (doi:10.1093/beheco/5.2.210)

299. Clayton DH. 1990 Mate choice in experimentally parasitized rock doves: lousy males lose. *Am Zool* **30**, 251–262. (doi:10.1093/icb/30.2.251)

300. Petrie M, Tim H, Carolyn S. 1991 Peahens prefer peacocks with elaborate trains. *Anim Behav* **41**, 323–331. (doi:10.1016/S0003-3472(05)80484-1)

301. Cherry MI. 1993 Sexual selection in the raucous toad, *Bufo rangeri*. *Anim Behav* **45**, 359–373. (doi:10.1006/anbe.1993.1041)

302. Senar JC, Negro JJ, Quesada J, Ruiz I, Garrido J. 2008 Two pieces of information in a single trait? The yellow breast of the great tit (*Parus major*) reflects both pigment acquisition and body condition. *Behaviour* **145**, 1195–1210. (doi:10.1163/156853908785387638)

303. Hõrak P, Ots I, Vellau H, Spottiswoode C, Pape Møller A. 2001 Carotenoid-based plumage coloration reflects hemoparasite infection and local survival in breeding great tits. *Oecologia* **126**, 166–173. (doi:10.1007/s004420000513)

304. Molnár O, Bajer K, Mészáros B, Török J, Herczeg G. 2013 Negative correlation between nuptial throat colour and blood parasite load in male European green lizards supports the Hamilton–Zuk hypothesis. *Naturwissenschaften* **100**, 551–558. (doi:10.1007/s00114-013-1051-4)

305. Safran RJ. 2004 Plumage coloration, not length or symmetry of tail-streamers, is a sexually selected trait in North American barn swallows. *Behav Ecol* **15**, 455–461. (doi:10.1093/beheco/arh035)

306. Hasegawa M, Arai E, Watanabe M, Nakamura M. 2014 Colourful males hold high quality territories but exhibit reduced paternal care in barn swallows. *Behaviour* **151**, 591–612. (doi:10.1163/1568539X-00003150)

307. Harper DGC. 1999 Feather mites, pectoral muscle condition, wing length and plumage coloration of passerines. *Anim Behav* **58**, 553–562. (doi:10.1006/anbe.1999.1154)

308. Pick JL, Nakagawa S, Noble DWAA. 2019 Reproducible, flexible and high-throughput data extraction from primary literature: the metaDigitise R package. *Methods Ecol Evol* **10**, 426–431. (doi:10.1111/2041-210X.13118)

309. Nakagawa S, Lagisz M, Jennions MD, Koricheva J, Noble DWA, Parker TH, Sánchez-Tójar A, Yang Y, O’Dea RE. 2021 Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods Ecol Evol* **2021**, 1–18. (doi:10.1111/2041-210X.13724)

310. R Core Team. 2024 R: a language and environment for statistical computing.

311. Viechtbauer W. 2010 Conducting meta-analyses in R with the metafor. *J Stat Softw* **36**, 1–48. (doi:10.18637/jss.v036.i03)

312. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis. *Evol Ecol* **26**, 1253–1274. (doi:10.1007/s10682-012-9555-5)

313. Paradis E, Schliep K. 2019 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. (doi:10.1093/bioinformatics/bty633)

314. Michonneau F, Brown JW, Winter DJ. 2016 rotl: an R package to interact with the Open Tree of Life data. *Methods Ecol Evol* **7**, 1476–1481. (doi:10.1111/2041-210X.12593)

315. OpenTreeOfLife, McTavish BR, Reyes LLS, Cranston KA, Allman J, Holder MT, McTavish EJ. 2019 Open Tree of Life Synthetic Tree. (doi:10.5281/zenodo.3937741)

316. Pollo P, Nakagawa S, Kasumovic MM. 2022 The better, the choosier: a meta-analysis on interindividual variation of male mate choice. *Ecol Lett* **25**, 1305–1322. (doi:10.1111/ele.13981)

317. Amrhein V, Greenland S, McShane B. 2019 Scientists rise up against statistical significance. *Nature* **567**, 305–307. (doi:10.1038/d41586-019-00857-9)

318. Hedges L V., Schauer JM. 2019 More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics* **44**, 543–570. (doi:10.3102/1076998619852953)

319. Schauer JM, Hedges L V. 2020 Assessing heterogeneity and power in replications of psychological experiments. *Psychol Bull* **146**, 701–719. (doi:10.1037/bul0000232)

320. Pottier P *et al.* 2024 Title, abstract and keywords: a practical guide to maximize the visibility and impact of academic papers. *Proc R Soc B* **291**. (doi:10.1098/rspb.2024.1222)

321. Foo YZ, O'Dea RE, Koricheva J, Nakagawa S, Lagisz M. 2021 A practical guide to question formation, systematic searching and study screening for literature reviews in

ecology and evolution. *Methods Ecol Evol* **12**, 1705–1720. (doi:10.1111/2041-210X.13654)

322. Schilhan L, Kaier C, Lackner K. 2021 Increasing visibility and discoverability of scholarly publications with academic search engine optimization. *Insights the UKSG journal* **34**, 1–16. (doi:10.1629/uksg.534)

323. Norry FM, Vilardi JC, Hasson E. 1997 Genetic and phenotypic correlations among size-related traits, and heritability variation between body parts in *Drosophila buzzatii*. *Genetica* **101**, 131–139. (doi:10.1023/A:1018360804439)

324. Seguin A, Forstmeier W. 2012 No band color effects on male courtship rate or body mass in the zebra finch: four experiments and a meta-analysis. *PLoS One* **7**, e37785. (doi:10.1371/journal.pone.0037785)

325. Boake CRB. 1985 Genetic consequences of mate choice: a quantitative genetic method for testing sexual selection theory. *Science* **227**, 1061–1063. (doi:10.1126/science.227.4690.1061)

326. Whittier TS, Kaneshiro KY. 1995 Intersexual selection in the Mediterranean fruit fly: does female choice enhance fitness? *Evolution* **49**, 990–996. (doi:10.1111/j.1558-5646.1995.tb02333.x)

327. Hedges L V., Olkin I. 1985 *Statistical methods for meta-analysis*. Amsterdam: Elsevier. (doi:10.1016/C2009-0-03396-0)

328. Jacobs P, Viechtbauer W. 2017 Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Res Synth Methods* **8**, 161–180. (doi:10.1002/jrsm.1218)

329. Cohen J. 1988 *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

330. Gurevitch J, Morrison JA, Hedges L V. 2000 The interaction between competition and predation: a meta-analysis of field experiments. *Am Nat* **155**, 435–453. (doi:10.1086/303337)

331. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. 1996 Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods* **1**, 170–177. (doi:10.1037/1082-989X.1.2.170)

332. Wan X, Wang W, Liu J, Tong T. 2014 Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol* **14**, 1–13. (doi:10.1186/1471-2288-14-135)

333. Nakagawa S, Cuthill IC. 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev* **82**, 591–605. (doi:10.1111/j.1469-185X.2007.00027.x)

Supplementary tables

Table S1. Proxies (i.e. variables related to condition-dependency, fitness, and other traits) considered valid in our study related to bearers of a sexual signal. *N* represents the number of meta-analytical studies in our dataset including each of these proxies. Underlined proxies are predicted to have a negative relationship with sexual signal expression, while others are predicted to have a positive relationship with sexual signal expression.

Proxy	**Sub proxy**	**Examples**
Attractiveness	Attractiveness or mating success (*N* = 15)	Copulation success, harem size, success in mate choice trials, divorce, pairing success or status
Benefits to sexual signal bearers or to mates	Latency (*N* = 9)	Latency to arrive at breeding site, latency to nest, latency to pair, latency to mate, latency to breed, latency to lay eggs, latency for eggs to hatch, latency for offspring to fledge
	Offspring quality or viability (*N* = 9)	Antioxidants or hormones in yolk, offspring growth rate, offspring attractiveness, offspring reproductive success, proportion of eggs hatched, fledging success, offspring size
	Parental care (*N* = 7)	Feeding rate, incubation frequency
	Paternity	Within and extra-pair paternity, cuckoldry occurrence

	(*N* = 10)	
	Reproductive success (*N* = 12)	Clutch size, breeding success, number of fledglings, total offspring sired
	Territory (*N* = 2)	Territory quality or size
Individual condition	Body condition (*N* = 11)	Carotenoid, protein, or lipid amount in plasma or in feathers, unspecified body condition, feather quality, subcutaneous fat score, residual mass, pectoral score
	External condition (*N* = 7)	Increase in brood size or reproductive effort, decrease in brood size or reproductive effort, habitat quality, mother's condition, date when reared, diet supplementation, dietary deprivation, nutritional stress
	Immune or antioxidant capacity (*N* = 10)	Antibody response, glucocorticoids, haematocrit, heterophil-to-lymphocyte ratio, oxidative damage, white blood cells
	Parasite resistance (*N* = 10)	Abundance of parasites, infection with a pathogen, pathogen richness, parasite removal
	Survival (*N* = 9)	Days alive, seen or re-captured after a given period
Other individual traits	Age (*N* = 8)	Age, ontogenetic stage (e.g. adult *vs*. juveniles)
	Aggression or social dominance (*N* = 7)	Performed aggression, received aggression, dominance, fights initiated, social rank, nest defence, distance from intruder, territory tenure
	Asymmetry (*N* = 3)	Fluctuating asymmetry
	Body size (*N* = 8)	Body (or part of it) mass, length, width, depth, area, or volume
	Brain size (*N* = 1)	Length, mass, area, or volume of the brain (or part of it)
	Heterozygosity (*N* = 2)	Multilocus heterozygosity, inbreeding
	Hormones (*N* = 2)	Androgens, estradiol, fertility phase, ovarian function, progesterone, testosterone
	Individual experience (*N* = 1)	Direct or indirect experience with individuals of the opposite sex (e.g. mated *vs*. unmated)
	Sexual signal (*N* = 4)	Another trait considered a sexual signal
	Traits related to sperm competition (*N* = 1)	Quantity of seminal fluid, sperm size, sperm viability, testes size

Table S2. Data points reported by meta-analyses that could not be re-extracted (*N* = 200). Reported IDs refer to the unique identifier for each reported data point across meta-analyses (see dataset).

Meta-analysis	**Reported IDs**	**Justification**
Dougherty 2021	106, 107	Reported data point indicates that both proxy and sexual signal (courtship behaviour) are from the male, but the latter was performed by the female (thus a measure of preference by females)

Dougherty et al. 2023	5028, 5046	Not found in the primary study: data were collected but were not reported
Dougherty et al. 2023	5064	Reported data point indicates that both proxy and sexual signal (courtship behaviour) are from the male, but the latter was performed by the female (thus a measure of preference by females)
Dougherty et al. 2023	5056, 5207	Considered repeated data: “brightness” is the average of other colour measures already reported
Evans et al. 2010	521, 522, 523	Not found in the primary study
Garamszegi & Eens 2004	4783, 4784, 4785, 4786, 4787, 4788, 4789, 4790, 4791	Not found in the primary study: data were collected but were not reported
Gontard-Danek & Moller 1999	648	Not found in the primary study: data were collected but were not reported
Griffith et al. 2006	727, 731	Considered repeated data: reported data point refers to a principal component measure as sexual signal that uses more granular data (hue, brightness, chroma)
Guindre-Parker & Love 2014	6780	Unable to extract this data point: slope without standard error
Hegyi et al. 2015	805, 807	Unable to extract this data point: slope without standard error
Hegyi et al. 2015	810	Not found in the primary study: data were collected but were not reported
Hernandez et al. 2021	4321	Not found in the primary study: data point seems to be related to arrival date instead of attractiveness and it is only approximately reported
Jennions et al. 2001	870	Not found in the primary study: only information given is on offspring survival not on survival of the sexual signal bearer
Mautz et al. 2013	969	Unable to extract this data point: Wald's Chi-Square without *p*-value
Mautz et al. 2013	994, 1009, 1014	Considered as repeated data: absolute number of live sperm is similar to data already given (number of sperm and percent of live sperm)
Meunier et al. 2011	1287, 1298, 1321	Not found in the primary study
Moller & Alatalo 1999	1423	Not found in the primary study: offspring size reported but not offspring survival
Moller & Jennions 2001	1512	Not found in the primary study: asymmetry not even mentioned
Moller & Jennions 2001	1599	Not found in the primary study: only relative parental care reported, thus not valid for extraction
Moller & Thornhill 1998	6696	Unable to extract this data point: Chi-Square from test with predictor variable with more than one level
Moller & Thornhill 1998	6728	Reported data point appears to refer to territory centrality, which we did not consider a measure of attractiveness
Moller & Thornhill 1998	6697	Not found in the primary study
Moore et al. 2016	1666, 1669	Not found in the primary study: data were collected but were not reported
Moore et al. 2016	1730, 1731	Not found in the primary study: data were collected but were not reported
Nakagawa et al. 2007	1804, 1805, 1806	Not found in the primary study: data appear to be related to the relationship between body index and dominance, not between bib size and body index as reported
Nakagawa et al. 2007	1801	Not found in the primary study: data were collected but were not reported

Nolazco et al. 2022	2805	Considered invalid for extraction: intra-copulation rate as proxy
Nolazco et al. 2022	2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821	Unable to extract this data point: slope without standard error
Nolazco et al. 2022	1991, 2222, 2503, 2504	Unable to extract this data point: slope without standard error
Nolazco et al. 2022	2091	Not found in the primary study
Nolazco et al. 2022	2337	Not found in the primary study
Nolazco et al. 2022	2287	Not found in the primary study: data were collected but were not reported
Nolazco et al. 2022	1853, 2363	Not found in the primary study: data were collected but were not reported
Parker 2013	6220, 6221	Data point referred to response between cage a and cage b, which were random regarding sexual signal and thus without biological meaning
Parker 2013	5945, 5946	Unable to extract this data point: *F*-value from test with predictor variable with more than one level
Parker 2013	5460, 5461, 5462	Not found in the primary study
Parker 2013	5573, 5574, 5575, 5576	Data point related to the interaction between testosterone and age on sexual signal, not only age
Parker 2013	5590, 5591, 5592	Data point related to the interaction between age and sexual signal on testosterone, not only sexual signal
Parker 2013	5594, 5595, 5596	Data point related to the interaction between age and sexual signal on corticosterone, not only sexual signal
Parker 2013	6127, 6128	Not enough details on diet and specimens used to be extracted
Parker 2013	6419, 6420, 6422, 6423, 6424, 6440, 6441, 6442, 6458, 6459, 6461, 6462, 6463, 6478, 6479, 6481	Data point related to the interaction between sex and diet on sexual signal, not only diet
Parker 2013	6085	Considered invalid for extraction: moult duration as proxy
Parker et al. 2006	2841, 2842	Unable to extract this data point: *U*-value without *p*-value
Parker et al. 2006	2868, 2869, 2870, 2871, 2872, 2874, 2875, 2877, 2878	Not found in the primary study: data were collected but were not reported
Prokop et al. 2012	3096, 3097, 3098, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107	Considered invalid for extraction: genetic correlation
Robinson & Creanza 2019	3227	Not found in the primary study: data were collected but were not reported
Robinson & Creanza 2019	3228	Unable to extract this data point: *F*-value from test with predictor variable with more than one level
Romano et al. 2017a	3397, 3399, 3401, 3403	Data point related to maternal care and male sexual signals, not female sexual signal
Romano et al. 2017a	3610, 3611, 3615, 3616	Not found in the primary study
Romano et al. 2017a	3480, 3481, 3485, 3488, 3489	Unable to extract this data point: slope without standard error
Romano et al. 2017a	3472, 3474	Unable to extract this data point: *F*-value from test with predictor variable with more than one level
Romano et al. 2017a	3478	Unable to extract this data point: Chi-Square from test with predictor variable with more than one level
Romano et al. 2017a	3529	Not found in the primary study: asymmetry not even mentioned

Romano et al. 2017a	3525, 3526, 3527, 3528	Unable to extract this data point: slope without standard error
Romano et al. 2017a	3469	Considered as repeated data: same information was extracted from subgroups
Romano et al. 2017a	3494, 3495, 3498, 3499, 3500	Unable to extract this data point: *F*-value from test with predictor variable with more than one level
Romano et al. 2017a	3519, 3520	Not found in the primary study
Romano et al. 2017a	3638, 3639	Not found in the primary study: data were collected but were not reported
Santos et al. 2011	3832	Not found in the primary study: only territory size and quality were shown but not aggressive interactions
Santos et al. 2011	3895	Not found in the primary study: territory defence was reported but not territory establishment
Simons & Verhulst 2011	3906	Not found in the primary study: only information on attractiveness given is related to diet not sexual signal
Simons et al. 2012	3997	Not found in the primary study: effect of tac only reported in relationship to treatments, no relationship between this measure and sexual signals
Simons et al. 2012	3965	Not found in the primary study: no mention whatsoever of PHA response assay in it
Simons et al. 2012	3920, 3929, 3983, 4058, 4059	Not found in the primary study: experiment involved removing parasites from one group and comparing with a control, all other proxies were related to this setting not to sexual signals
Simons et al. 2012	3956	Not found in the primary study: only plumage colour reported
Soma & Garamszegi 2011	6547	Unable to extract this data point: *F*-value from test with predictor variable with more than one level
Street et al. 2016	4108, 4114, 4120, 4122, 4124, 4126	Not found in the primary study: data were collected but were not reported
Thornhill & Moller 1998	6666	Unable to extract this data point: Chi-Square from test with predictor variable with more than one level
Thornhill & Moller 1998	6669	Not found in the primary study: data were collected but were not reported
Thornhill & Moller 1998	6667, 6668	Not found in the primary study
Weaver et al. 2018	4526, 4531	Considered as repeated data: same information was extracted from subgroups
Weaver et al. 2018	4529	Considered invalid for extraction: proxy is from both males and females while sexual signal is only from males
White 2020	4555	Not found in the primary study: data were collected but were not reported
White 2020	4685, 4686, 4687, 4688, 4689	Not found in the primary study: species reported in meta-analysis is not even mentioned
White 2020	4690, 4691	Not found in the primary study: no parasite is mentioned
Yasukawa et al. 2010	4750, 4751	The primary study provided the difference in sexual signal between recaptured and non-recaptured individuals, which was considered a measure of survival in this meta-analysis. Yet, we disagree with this interpretation as the authors of the primary study do not suggest this is the case and the time between capture and recapture was short
Yasukawa et al. 2010	4721, 4722, 4723, 4740, 4741	Not found in the primary study: other measures of aggression given, but not dominance rank

Table S3. Evaluation of publication bias (small-study effect) of meta-analyses reported results. We considered that publication bias was detected (i.e. TRUE, shaded rows) when the slope 95% confidence interval was positive.

Meta-analysis	**Slope**	**Standard error**	**95% CI upper bound**	**95% CI lower bound**	**Publication bias detected**	**Publication bias reported**
Dougherty 2021	0.046	0.024	-0.001	0.092	FALSE	undetected
Dougherty et al. 2023	0.003	0.026	-0.048	0.054	FALSE	undetected
Evans et al. 2010	0.048	0.060	-0.070	0.166	FALSE	not assessed
Fiske et al. 1998	0.176	0.081	0.017	0.335	TRUE	not assessed
Garamszegi & Eens 2004	0.029	0.073	-0.114	0.171	FALSE	undetected
Garamszegi & Moller 2004	-0.186	0.257	-0.690	0.318	FALSE	not assessed
Garamszegi 2005	-0.017	0.151	-0.314	0.279	FALSE	not assessed
Garamszegi et al. 2007	0.605	0.247	0.122	1.089	TRUE	not assessed
Gontard-Danek & Moller 1999	0.138	0.097	-0.052	0.328	FALSE	undetected
Griffith et al. 2006	0.121	0.049	0.025	0.217	TRUE	detected
Guindre-Parker & Love 2014	0.164	0.053	0.060	0.269	TRUE	detected
Hegyi et al. 2015	-0.005	0.035	-0.073	0.064	FALSE	undetected
Hernandez et al. 2021	0.085	0.039	0.008	0.161	TRUE	detected
Jennions et al. 2001	0.093	0.033	0.027	0.158	TRUE	detected
Koch et al. 2016	-0.160	0.156	-0.465	0.145	FALSE	undetected
Mautz et al. 2013	-0.098	0.038	-0.173	-0.023	FALSE	undetected
McLean et al. 2012	-0.093	0.108	-0.305	0.119	FALSE	not assessed
Meunier et al. 2011	0.078	0.053	-0.025	0.181	FALSE	undetected
Moller & Alatalo 1999	0.054	0.091	-0.125	0.233	FALSE	not assessed
Moller & Jennions 2001	0.078	0.105	-0.127	0.283	FALSE	undetected
Moller & Thornhill 1998	-0.083	0.060	-0.199	0.034	FALSE	not assessed
Moore et al. 2016	0.006	0.043	-0.077	0.090	FALSE	undetected
Nakagawa et al. 2007	0.106	0.148	-0.184	0.396	FALSE	undetected
Nolazco et al. 2022	0.034	0.018	-0.003	0.070	FALSE	undetected
Parker 2013	0.097	0.035	0.029	0.165	TRUE	detected
Parker et al. 2006	0.202	0.127	-0.048	0.451	FALSE	detected
Prokop et al. 2012	-0.010	0.059	-0.126	0.106	FALSE	undetected
Robinson & Creanza 2019	0.091	0.055	-0.017	0.198	FALSE	undetected
Romano et al. 2017a	0.011	0.019	-0.026	0.048	FALSE	undetected
Romano et al. 2017b	0.043	0.054	-0.063	0.148	FALSE	undetected
Sanchez-Tojar et al. 2018	-0.067	0.102	-0.267	0.133	FALSE	undetected
Santos et al. 2011	0.067	0.039	-0.009	0.142	FALSE	undetected
Simons & Verhulst 2011	-0.038	0.160	-0.353	0.276	FALSE	undetected
Simons et al. 2012	0.052	0.036	-0.020	0.123	FALSE	undetected
Soma & Garamszegi 2011	0.190	0.038	0.115	0.264	TRUE	detected
Street et al. 2016	-0.150	0.156	-0.456	0.157	FALSE	not assessed
Weaver et al. 2018	0.035	0.025	-0.014	0.084	FALSE	undetected
White 2020	0.055	0.033	-0.010	0.120	FALSE	undetected
Yasukawa et al. 2010	0.165	0.074	0.021	0.309	TRUE	not assessed

Table S4. Slopes from linear regressions between originally reported effect sizes by meta-analyses and effect sizes that were re-extracted from their sources (but only those that exactly matched their description). Shaded rows highlight slopes whose 95% confidence interval does not include 1.

Meta-analysis	**Slope**	**Standard error**	**95% CI upper bound**	**95% CI lower bound**
Dougherty 2021	0.976	0.101	1.193	0.759
Dougherty et al. 2023	0.956	0.037	1.029	0.883
Fiske et al. 1998	0.254	0.150	0.620	-0.112
Garamszegi & Eens 2004	1.008	0.019	1.047	0.970
Garamszegi & Moller 2004	0.642	0.138	0.981	0.304
Gontard-Danek & Moller 1999	0.828	0.218	1.388	0.268
Griffith et al. 2006	0.696	0.119	0.939	0.453
Hegyi et al. 2015	0.720	0.050	0.839	0.601
Hernandez et al. 2021	0.022	0.132	0.285	-0.241
Jennions et al. 2001	1.026	0.315	1.720	0.332
Mautz et al. 2013	0.619	0.119	0.859	0.380
McLean et al. 2012	0.778	0.055	0.906	0.651
Meunier et al. 2011	0.743	0.248	1.274	0.212
Moller & Jennions 2001	0.311	0.117	0.569	0.053
Moore et al. 2016	0.371	0.311	1.171	-0.429
Nakagawa et al. 2007	0.907	0.043	0.994	0.819
Nolazco et al. 2022	0.755	0.031	0.816	0.693
Parker & Ligon 2003	1.047	0.108	1.311	0.783
Parker 2013	0.460	0.061	0.581	0.339
Parker et al. 2006	0.984	0.011	1.007	0.962
Robinson & Creanza 2019	0.876	0.082	1.043	0.709
Romano et al. 2017a	0.316	0.056	0.428	0.204
Romano et al. 2017b	0.996	0.018	1.036	0.955
Sanchez-Tojar et al. 2018	0.868	0.355	1.736	0.001
Santos et al. 2011	0.800	0.152	1.127	0.472
Simons et al. 2012	0.729	0.087	0.907	0.552
Soma & Garamszegi 2011	0.977	0.041	1.059	0.895
Street et al. 2016	0.123	0.077	0.286	-0.041
Weaver et al. 2018	0.663	0.069	0.800	0.527
White 2020	0.760	0.082	0.929	0.592

Table S5. Recommendations for meta-analytical research projects.

Item	**Recommendation(s)**
Data of interest	• Detail all criteria and data as much as possible instead of mentioning umbrella terms • Provide data details in a readable format, not only within the dataset
Direction and extraction of effect sizes	• Establish a coherent rationale to ascertain direction of effect sizes, fully reporting it in the manuscript • Double-check extracted data (done by a researcher other than the data extractor)

- Provide the within-text location of data extracted
- Provide equations used to calculate effect sizes, along with assumptions and data transformations utilised

Supplementary figures

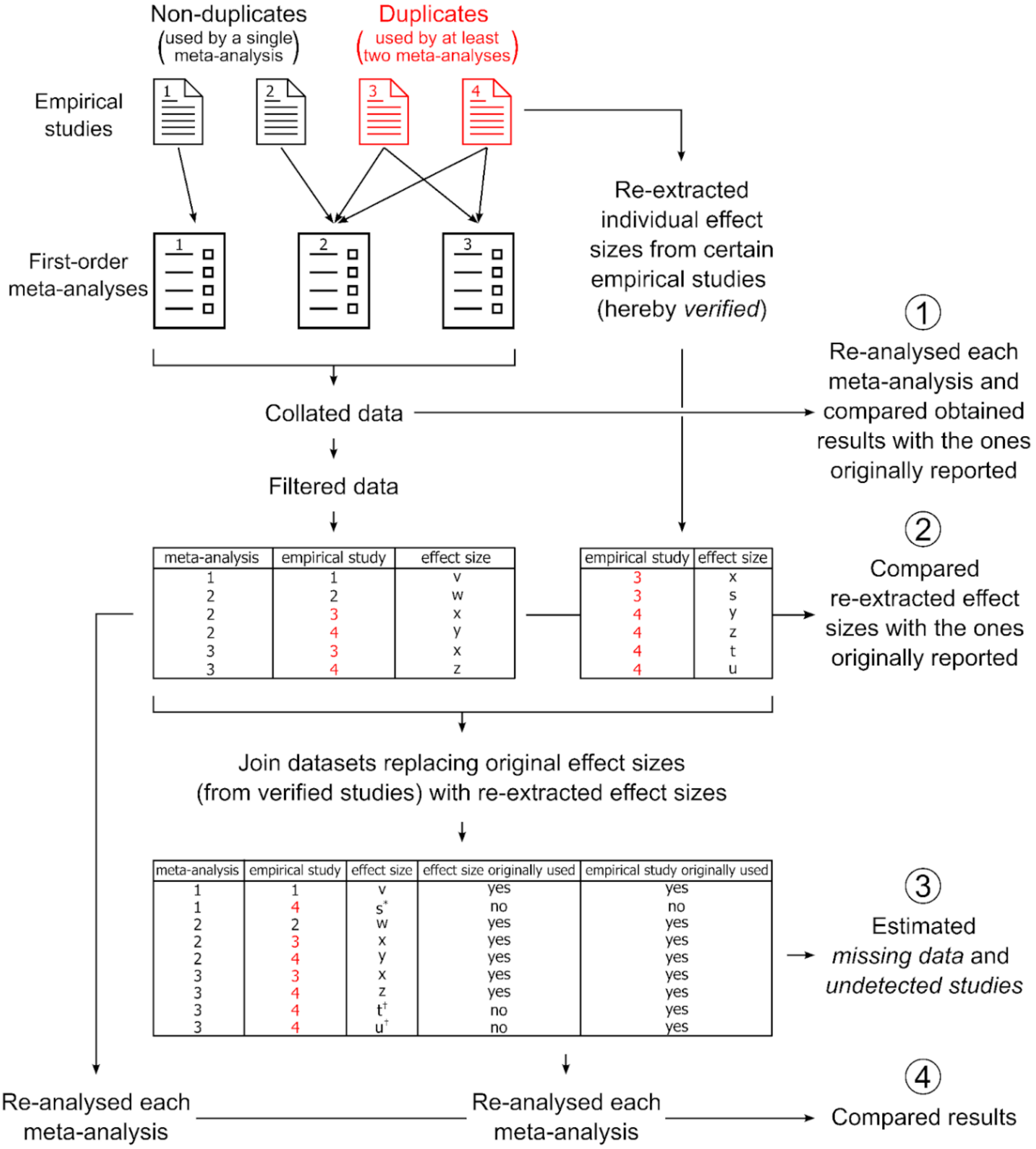

Figure S1. Summarised workflow used in our study, with four main aims. *Examples of effect sizes that were not reported in a meta-analysis because the empirical study containing it was not originally detected (i.e. *undetected studies*). †Examples of effect sizes that were not originally reported by a meta-analysis even though the empirical study was listed as a source for other effect sizes (i.e. *missed data*).

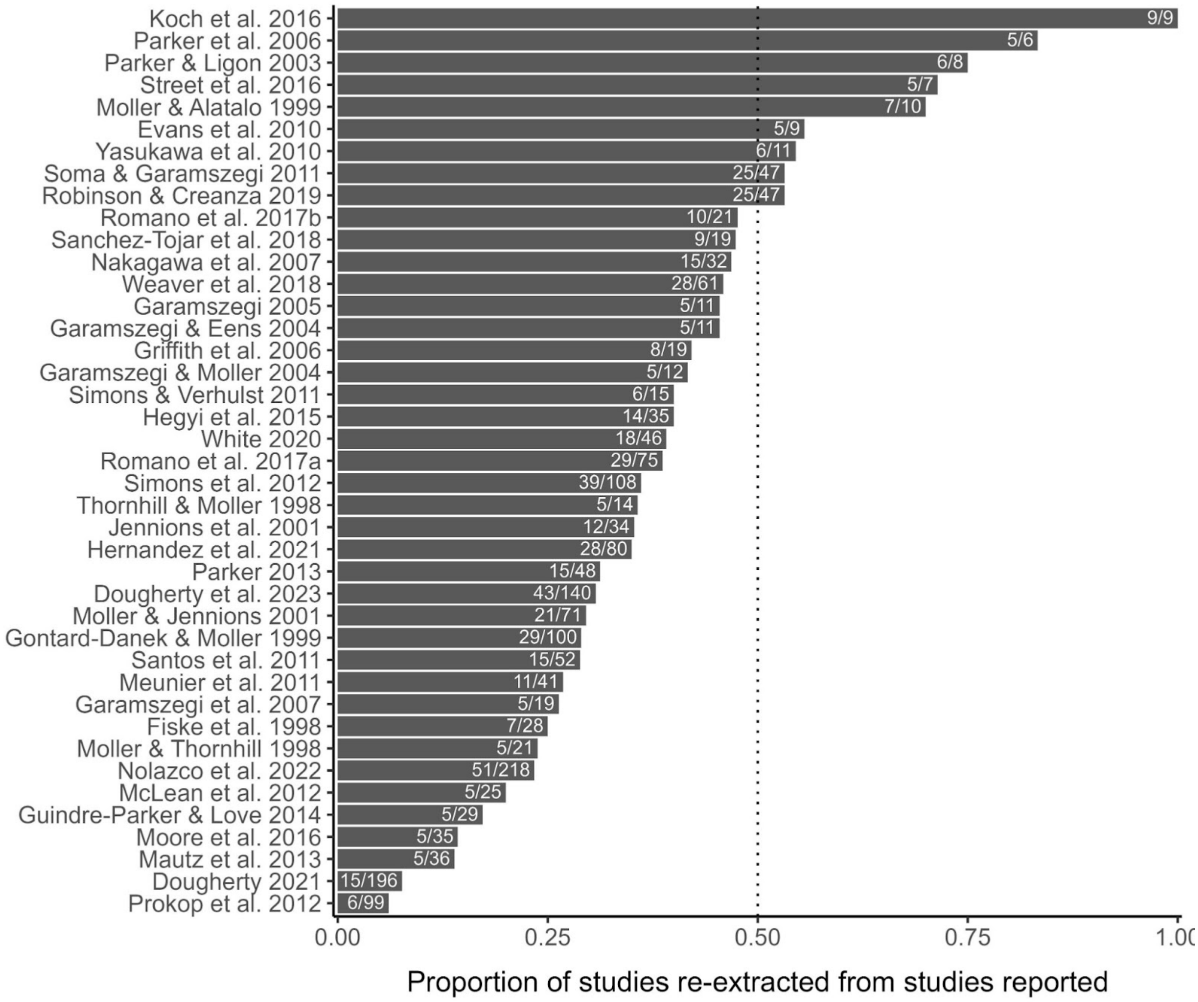

Figure S2. Proportion of empirical studies that we verified (i.e. from which we re-extracted data) from the number of empirical studies reported as sources by each meta-analysis. The number of empirical studies we verified out of the number of empirical studies reported as sources by each meta-analysis is shown within bars. The dotted line highlights when half of the empirical studies were verified.

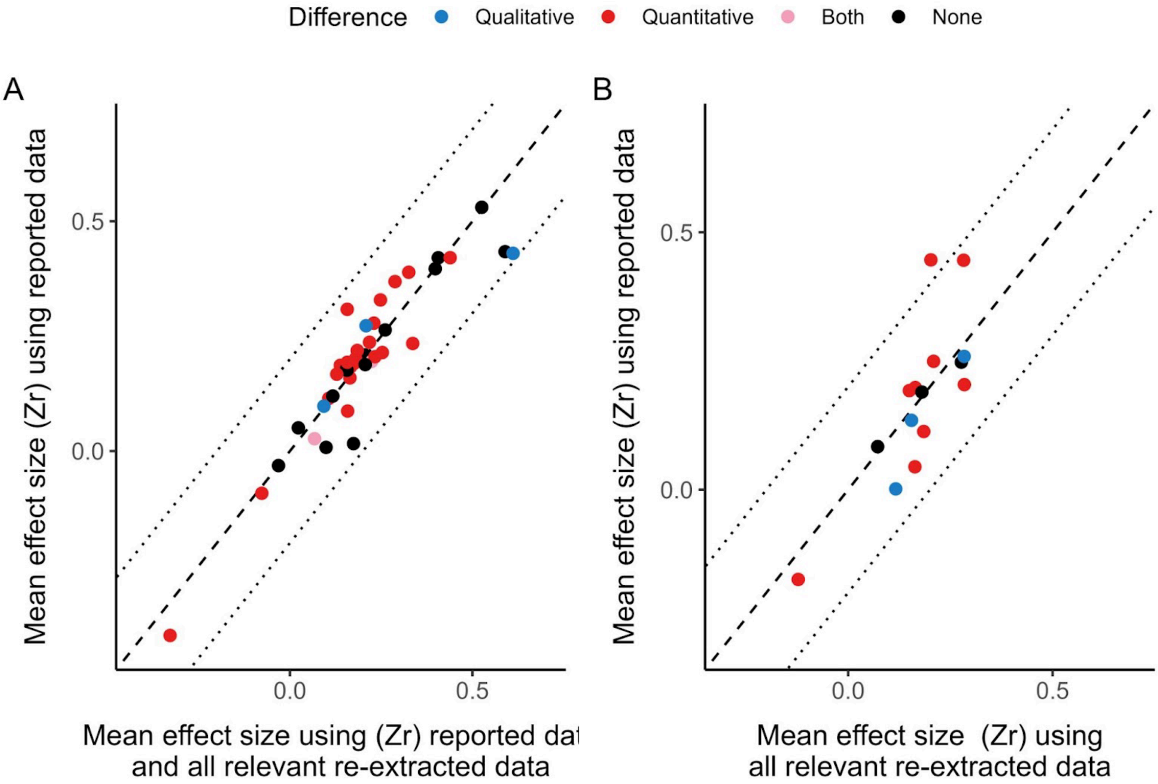

Figure S3. Comparison between mean effect sizes resulting from analyses using the dataset originally reported by meta-analyses and those resulting from analyses incorporating all relevant re-extracted data (matched, missed, and undetected). Data from both verified and unverified empirical studies are analysed in A, while only data from verified studies are used in B. Dashed lines highlight perfect correspondences between variables, while dotted lines highlight a difference of 0.2 and -0.2 between variables.

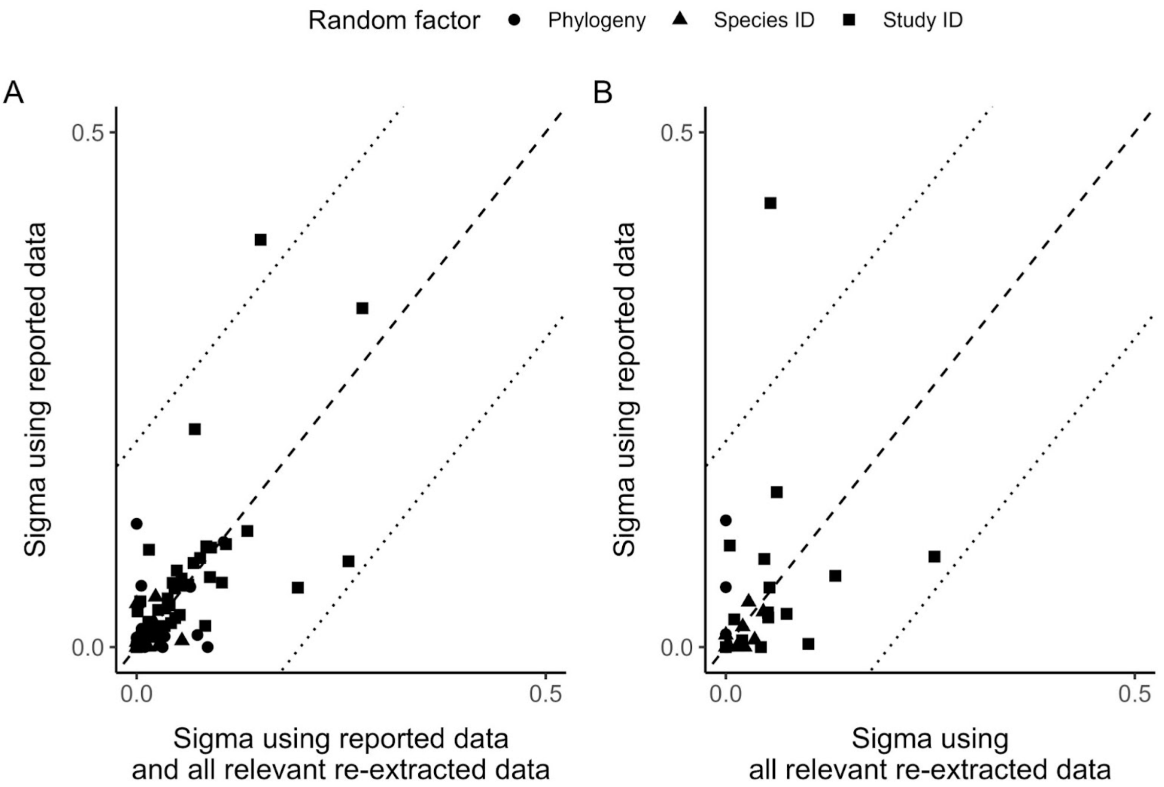

Figure S4. Comparison between sigmas (from each random factor) resulting from analyses using the dataset originally reported by meta-analyses and those resulting from analyses incorporating all relevant re-extracted data (matched, missed, and undetected). Data from both verified and unverified studies are analysed in A, while only data from verified empirical studies are used in B. Dashed lines highlight perfect correspondences between variables, while dotted lines highlight a difference of 0.2 and -0.2 between variables.

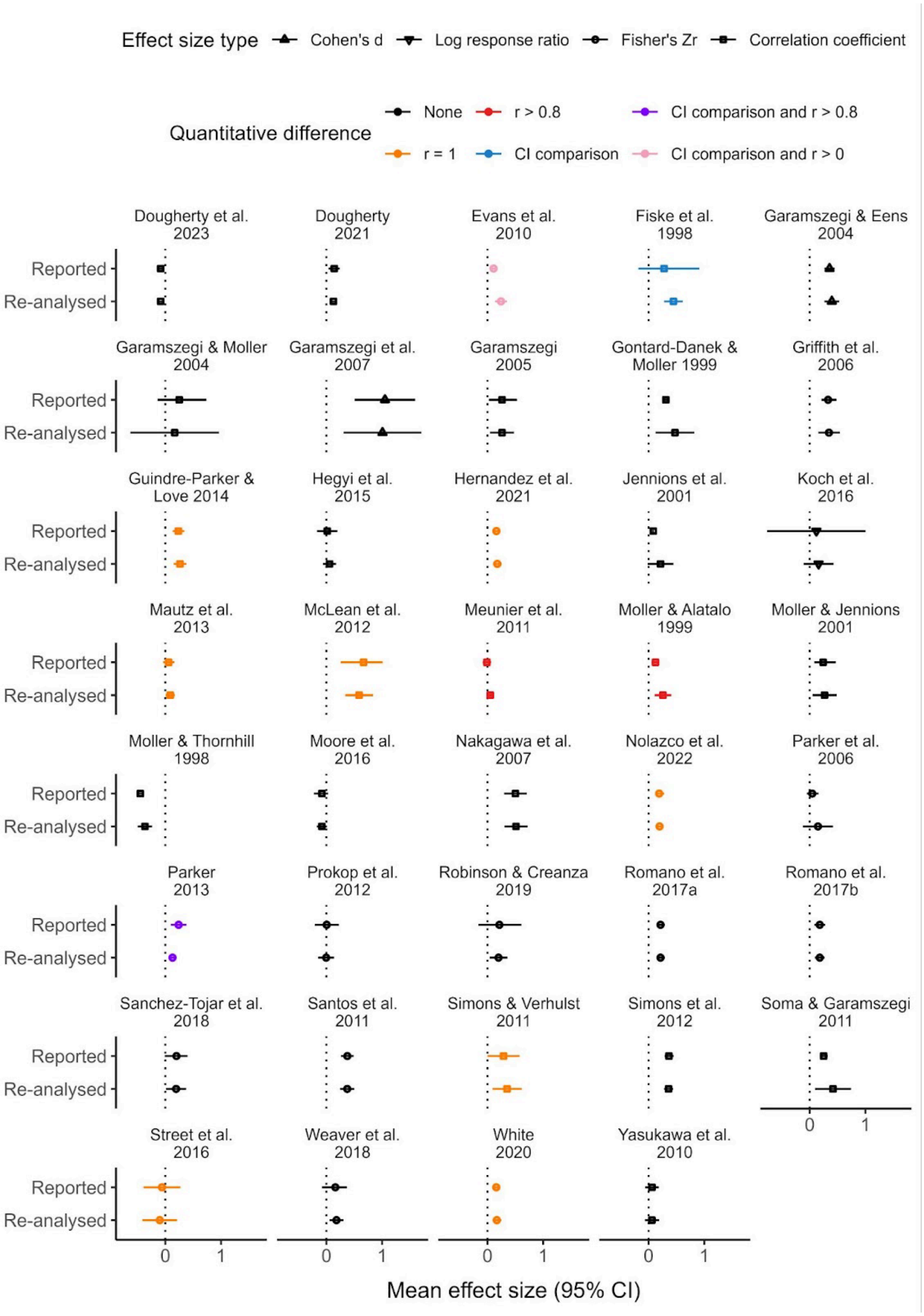

Figure S5. Originally reported and re-analysed mean effect sizes of 39 meta-analyses. Colourful points indicate quantitative pairwise differences with absolute *z*-score greater than

1.96 (using different values for r) and/or when the 95% confidence interval of the re-analysed mean effect size did not include the value of the reported mean effect size (“CI comparison”).

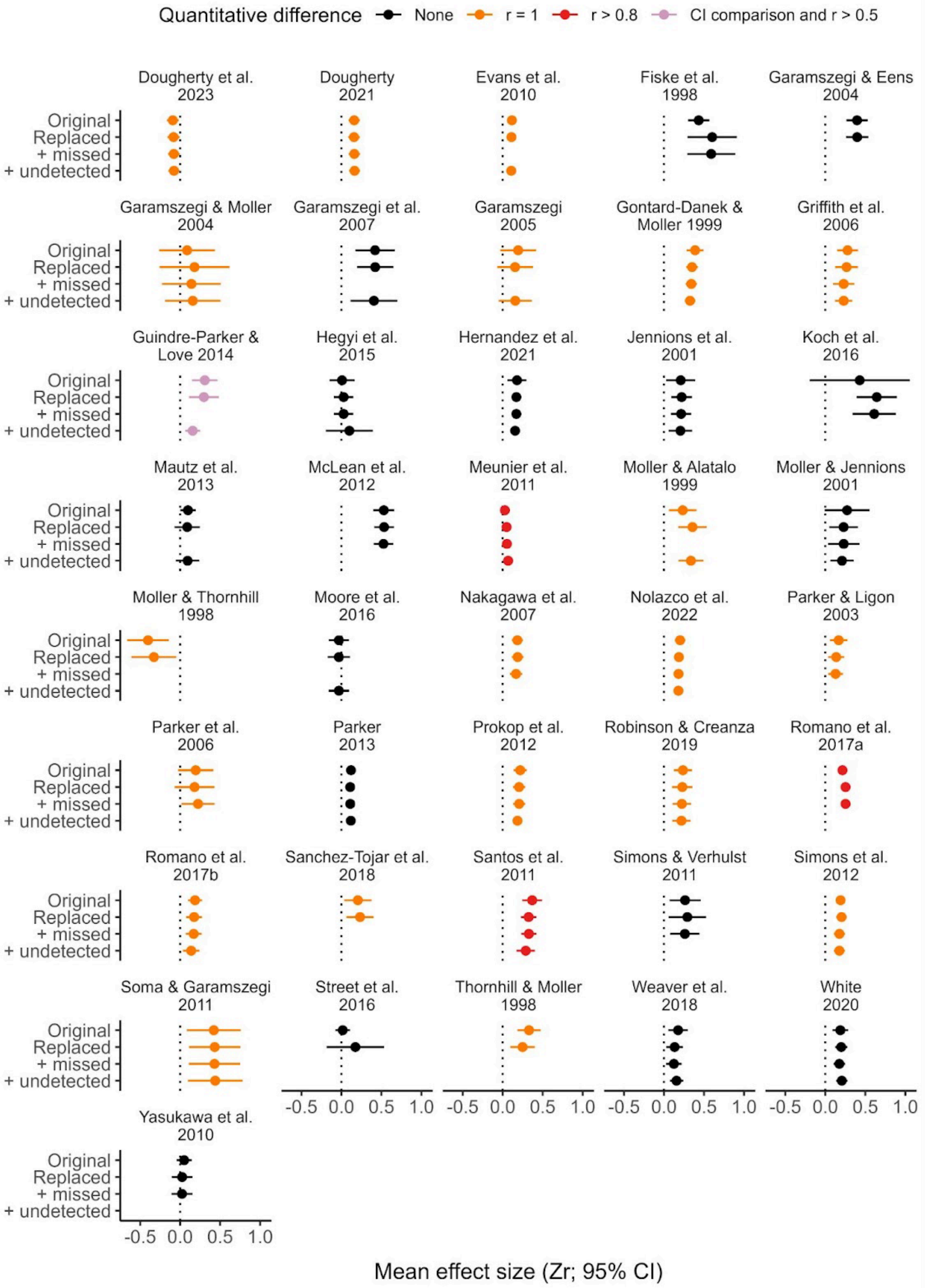

Figure S6. Mean effect sizes from up to four distinct meta-analytical models for each of 41 meta-analyses, using data from all empirical studies (both verified and unverified by us, see details in-text). Dotted lines highlight zero. Colourful points indicate quantitative pairwise differences (between the first and the last result shown within each subplot) with absolute *z*-score greater than 1.96 (using different values for *r*) and/or when the 95% confidence interval of the re-analysed mean effect size did not include the value of the reported mean effect size ("CI comparison").

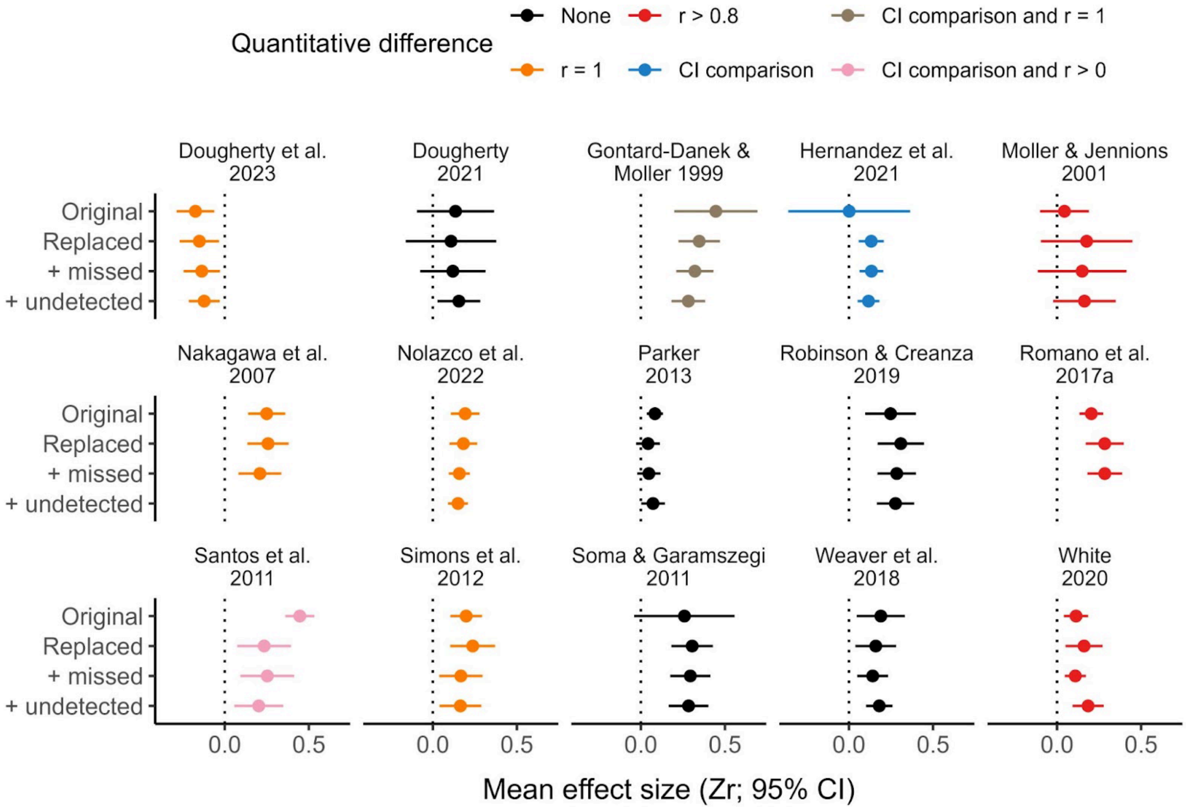

Figure S7. Mean effect sizes from up to four distinct meta-analytical models for each of 15 meta-analyses, using data from verified empirical studies (see details in-text). Dotted lines highlight zero. Colourful points indicate quantitative pairwise differences (between the first and the last result shown within each subplot) with absolute *z*-score greater than 1.96 (using different values for *r*) and/or when the 95% confidence interval of the re-analysed mean effect size did not include the value of the reported mean effect size ("CI comparison").

Supplementary information S1. Changes from the pre-registration

We replaced terms used in our pre-registration regarding our goals (e.g. “precision”, “completeness”, “detectability”) with other terms (e.g. number of undetected studies) to improve the clarity of our manuscript. Furthermore, we created different proxy categories (e.g. “heterozygosity”, “hormones”, “territory”) and reorganised proxies’ master categories to improve clarity and better show their diversity (see Table S1).

Supplementary information S2. Data of interest

First, for the purpose of our study, we interpret ornaments and sexual signals as “non-ordinary and non-weapon traits potentially used for mate attraction”. This vague definition is intended to match the rationale observed in the literature, in which conspicuous and/or sexually dimorphic traits are often subjectively designated as sexual signals (see also the discussion in [7]). In practice, this means that we included all traits considered as sexual signals by meta-analyses’ authors, except for obvious weapons (i.e. antlers, spurs, and horns), ordinary traits (i.e. body size; body parts without referring to it as a sexual trait, e.g. [323]), lek related measures (e.g. lek attendance, distance from lek centre, etc.; see [17]) colour bands put on animals by researchers (see [324]), and comparisons between “attractive” *vs*. “unattractive” based on mate preferences without a direct measurement of a sexual signal (e.g. [325,326]). In addition, we only considered asymmetry as a sexual signal when it was related to a trait that itself was considered a sexual signal. The only extended phenotypes we considered as valid sexual signals were measures of structures to attract mates (e.g. bowers, nests). This means that data related to leks (e.g. lek attendance, distance to centre of the lek) and territory quality were not considered sexual signals. Second, we are interested in how sexual signals relate to distinct conditions, fitness proxies, and individual traits (Table S1). However, we excluded effect sizes related to traits of focal individuals’ mates (e.g. parental care provided by mate), assortative mating, heritability (i.e. proportion of phenotypic variance explained by genetic variance), and offspring sex ratio because we deemed these estimates may not be linked to benefits to or reflect patterns of sexual signal bearers. Third, we excluded effect sizes related to humans to avoid confounding our cultural influences with biological aspects of non-human animals (see also [7]). Fourth, we excluded effect sizes with (exact) zero values from three meta-analyses that purposefully included non-significant results as such [14,15,39], unless we were able to match them with at least one re-extracted data point (see *Re-extractions* section).

Supplementary information S3. Priority and selection of empirical studies for data re-extraction

We classified empirical studies (used as sources for data extraction in meta-analyses) according to the following categories. We first classified all empirical studies listed as sources for at least three meta-analyses as "high priority" (n = 82). Then, we classified the remaining duplicates (i.e. empirical studies that were used exactly by two meta-analyses) as: (1) "low priority" when empirical studies whose all effect sizes were identical (both in their description and in their value) across meta-analyses (n = 19); (2) "medium priority" when empirical studies contained effect sizes regarding different traits and measurements (e.g. distinct sexual signals or fitness proxies) across meta-analyses (n = 79); (3) "high priority" when empirical studies generated similarly described effect sizes (e.g. same sexual signal and fitness proxy) but with different (and thus contradicting) values or sample sizes across meta-analyses (n = 136). We primarily targeted high priority duplicates (n = 218) for effect size re-extraction, although not all meta-analyses contained these duplicates (e.g. [18,50]). We thus established a minimum of five empirical studies for re-extraction of effect sizes per meta-analysis, adding medium priority duplicates (n = 12) and non-duplicated empirical studies (n = 19) when necessary to meet this quota.

Supplementary information S4. Meta-analyses' inclusion criteria

Below, we provide excerpts from the meta-analyses in our dataset related to their data of interest and inclusion criteria. We also provide the issues we detected for each meta-analysis (if any).

Dougherty (2021)

<u>Originally reported</u>

"I focused on sexual signalling behaviour, including long-range attraction signals and short-range courtship signals. I excluded measures of investment in non-behavioural signals, such as ornaments or advertising colours, although I did not exclude display behaviours that involved such ornaments. I excluded intrasexual signals or signals for which a primary intrasexual function could not be ruled out. However, I acknowledge that all sexual signals probably function intrasexually to some extent. I also included several lekking species for which displays probably signal to potential mates and rivals because I consider the primary function of leks to be mate assessment. I included studies examining both male and female sexual signalling. I included acoustic, visual, olfactory (pheromone) and tactile display behaviours. I focused on behavioural traits that reflect the motivation to signal (for example, courtship latency) or energetic investment in signalling (such as signalling duration, rate and some measures of intensity). For acoustic signals, I included measures of call loudness, except when related to body size (because call loudness may be constrained by the size of the sound-producing organs). I excluded measures of signal complexity because this does not necessarily relate to overall energetic investment per se. For acoustic signals, I excluded measurements of call pitch/frequency and fine-scale temporal components of a call. For pheromones, I excluded measurements of pheromone composition but included measures of time spent releasing pheromone and pheromone titre if measured outside the body (I excluded measures of pheromone titre in dissected glands or bodies).

I included studies examining how sexual signalling behaviour related to age. I included studies in this category if all individuals were sexually mature and age was not confounded with body size. Importantly, age is often confounded by mated status, especially for wild individuals and this may influence the motivation to signal independently of age. Therefore, I only considered studies examining age-related signalling in virgins. This was necessary because few studies record both age and mated status in a way that allows their independent effects to be estimated. [...] I included studies examining how sexual signalling behaviour related to other measures of attractiveness. In these studies, attractiveness was determined in relation to behavioural or morphological traits which were suggested to signal mate quality (either genetic quality or current condition) or have been shown to be used in mate choice. I included studies relating signalling behaviour to: (1) song quality, (2) ornament size, (3) ornament or body colouration or brightness, (4) morphological asymmetry, (5) inbreeding, (6) territory or nest quality and (7) social rank. Individuals were assumed to be attractive if they exhibited high-quality song, large ornaments, bright or intense colouration, were outbred, with low asymmetry, of a high social rank and had high-quality territories or nests. I included tests of social rank only when signalling was recorded in the absence of rivals; this is important because high-ranking individuals may suppress the behaviour of subordinates.

This category included studies relating signalling to body length, weight or some proxy length measurement (for example, leg length, wing length and pronotum width).

I included studies examining how sexual signalling behaviour related to several measures of condition: (1) diet or food level, (2) the relationship between body size and weight, (3) direct measurements of body lipid content or plasma metabolite level and (4) environmental conditions that could alter physiological stress in the short-term (oxygen, carbon dioxide and water acidity in aquatic environments). I used several indirect, morphological measures of condition, although I note that several common measures have been criticised. I assume that individuals were in poor condition if they were relatively light for a given body size, with low lipid stores, or had experienced low food levels, poor-quality diets or stressful environments. I excluded studies examining how signalling behaviour related to physiological markers of stress, as stress responses are typically short-lived and may have a complex relationship with condition.

I included studies comparing signalling between virgins and once-mated individuals. I excluded tests related to the number of matings above one or other forms of mating experience (that is, the phenotype of previous mating partners).

I included studies examining how sexual signalling behaviour related to parasite load. This category included studies measuring the presence or number of external (lice, mites and crustaceans) or internal (acanthocephalans, nematodes, platyhelminthes, alveolates, fungi, bacteria and viruses) parasites. I included sexually transmitted parasites, even in cases where host behavioural changes were suggested to be due to parasite manipulation. I excluded studies relating behaviour to the presence of endosymbionts in insects, as they may have parasitic or mutualistic effects on their hosts. Finally, I also excluded studies examining the effect of controlled immune challenges on host behaviour; for example, by introducing sterile pellets or inactivated pathogens into the host. This is because any consequences for host condition are indirect in such cases, caused by upregulation of the host immune system and are typically short-lived."

<u>Issues</u>

1) The dataset contains sexual signals that are behavioural products (i.e. extended phenotype; e.g. number of pillars built, domes built, and building rate in different species of crabs). The

manuscript does not mention this or other extended phenotypes. We thus considered data on extended phenotype as valid sexual signals for matching purposes.

2) The manuscript does not mention brood size manipulation as a valid proxy even though several data points related to this proxy are in the reported dataset. We thus considered brood size manipulation as a valid proxy for matching purposes.

Dougherty et al. (2023)

Originally reported

“To be included in the analysis, a study had to (a) present data for sexually mature individuals of a non-human animal species, (b) report within-species variation in a morphological, behavioural, or extended ornament, (c) report some measure of parasite load for the same host individuals, and (d) provide sufficient statistical information for an effect size to be calculated. While parasite-mediated sexual selection hypotheses tend to focus on elaborate male morphological ornaments, such as plumage or bright skin patches, we expand the scope of our data set by (a) considering display behaviours and extended ornaments such as the bowers of bowerbirds because these potentially honestly indicate courter condition or quality and (b) considering female ornamentation because mating preferences in relation to partner condition and quality are seen in both sexes.”

Issues

1) Many aspects from acoustic signals (e.g. repertoire size, song switching rate) seemed to have been excluded without mentioning this in-text, which we interpreted as a hidden criterion.

2) The dataset contains details only for some sexual signals, while others remain unclear (e.g. "colour" without specifying body part or how it was measured).

Evans et al. (2010)

Originally reported

“We conducted a search of peer-reviewed journals to identify studies that examined ventral plumage reflectance in free-living great tit populations and also contacted known researchers directly to ask them to contribute data. [...] The data in the meta-analysis included measures derived from spectrometry, colorimetry, digital photography, and colour tiles, and were based on measures of live birds and of sampled feathers. As a result of the limited number of studies represented, we did not test for systematic differences with respect to methodology. We did not impose restrictions on the colour measures that would be included in our meta-analysis; rather, we invited researchers to use colour measures of their choice. [...] To examine the overall extent of sex- and age-dependent colour expression, we compiled data on the mean, standard deviation and sample size for each sex (males versus females) and, separately, each age class (first-years versus older birds).”

Issues

No issues detected.

Fiske et al. (1998)

Originally reported

“We included studies if they contained correlates (parametric Pearson or nonparametric Spearman correlation coefficients) between traits and male mating success. Assessing male mating success across several taxa is not an easy task. In many lekking species (e.g., certain insects) copulations are hard or even impossible to observe. Two of the studies included presented their results as correlates of female visitation rates. This is probably legitimate because this measure is highly intercorrelated with the number of copulations. Still, our analyses were mainly based on studies that obtained correlations with male copulatory success. This led to rejection of some ungulate studies because they reported correlations with territory mating success. [...] Attendance: We included studies that provided information about the time individual males spent on the lek in relation to male mating success. This variable was measured both as the proportion of time compared to other males and as absolute time within a given period. Display frequency: Males vocalisations are probably important traits when females make their mate choice decision. Therefore we included measurements of display frequencies (calls per unit time or proportions of time calling) because such measurements can easily be compared across species. Spectral aspects of calls have also been analysed in several studies, but we have chosen not to include them in our analyses because we found them impossible to categorise uniformly."

Issues

1) All data reported were related to mating success, but authors never provided the specific measure for each data point or study that represented this proxy. We thus accepted all proxies of mating success for matching purposes.

Garamszegi (2005)

Originally reported

"I performed some simple meta-analyses on the available data to test for general patterns in the intraspecific association between parasitism and bird song. [...] I used measures of immune function and parasite prevalence interchangeably, as was done in similar meta-analyses."

Issues

No issues detected.

Garamszegi & Eens (2004)

Originally reported

“We collected published results of studies investigating correlatively the intraspecific association between male repertoire size and/or song length and the size of HVC and/or RA. Hence, we did not include studies that investigated seasonal variation in RA and HVC volumes and song by sampling males in different seasons, or compared groups of birds experiencing different tutoring regimes or originating from different populations. [...] As some studies distinguished between the absolute and relative volumes of brain nuclei by taking or not the covariation with telencephalon into account, we also estimated effect sizes for absolute and relative volumes of the HVC and RA.”

Issues

No issues detected.

Garamszegi & Moller (2004)

Originally reported

“We collected published and, in an effort to control for publication bias, unpublished results of studies investigating the association between songs and extrapair paternity within species. [...] We included analyses from t-tests (two-tailed), or other equivalent statistics testing the null hypothesis that males with more extravagant song display (measured as repertoire size, song rate, song length, and performance characteristics) have similar paternity in their own nests than do males with less extravagant song features.”

Issues

1) Manuscript is unclear on which measures of paternity were included, so we considered all extra- and within-paternity measures as valid for matching purposes.

2) Manuscript is unclear on which song traits were included while the dataset contains various song traits. We thus considered all possible song traits as valid for matching purposes.

Garamszegi et al. (2007)

Originally reported

“[...] we collected published results of studies investigating the relationship between age and song using cross-sectional approaches. We included analyses based on t-test (two-tailed), or other equivalent statistics testing the null hypothesis that yearling males have similar repertoire to older males.”

Issues

No issues detected.

Gontard-danek and Moller (1999)

Originally reported

“We searched the literature for correlation coefficients, or other statistics that could be converted into correlation coefficients, based on the relationship between sexual selection and the expression of secondary sexual characters. […] We considered whether effect size differed in relation to the currency used to estimate success. These were classified as mate preference, mating success, breeding success, or paternity.”

Issues

1) The manuscript mentions proxies without details ("mate preference, mating success, breeding success, or paternity"). We thus considered all measurements related to these proxies (except those related to timing or latency) as valid for matching purposes.

2) The dataset contains one data point related to territory quality even though the manuscript only mentioned other proxies. We thus included relevant data that used territory quality as proxies.

3) The manuscript only mentions "secondary sexual characters", which apparently excludes behavioural sexual signals as these were not present in the dataset. We thus interpreted this as a hidden criterion and did not consider behavioural sexual signals as valid for matching purposes.

4) Dataset only contained measurements on timing or latency related to breeding, egg laying, mating, and pairing. We thus interpreted this an omitted criterion and excluded other similar measures (e.g. arrival date) for matching purposes.

Griffith et al. (2016)

<u>Originally reported</u>

“A number of experimental manipulations have been conducted to test the condition dependence of carotenoid- and melanin-pigmented ornaments, especially in birds. These studies manipulated condition-related factors including diet, parasite load and parental effort. We did not include studies that manipulated the carotenoid content of the diet because the absence of carotenoids in the diet must lead to the absence of carotenoids in the body, and thus a response to this treatment does not demonstrate condition dependence of a carotenoid-pigmented trait.

Pigmented ornaments were assessed for changes in area, brightness, hue, saturation, chroma, principal component-defined colour variation, or other colour score. We attempted to locate all such studies from birds for inclusion in meta-analyses. All studies included in the meta-analyses examined the response of traits that appear ornamental and thus may have evolved in response to sexual selection."

<u>Issues</u>

1) The dataset contains a data point related to assessment of a sexual signal before and after an immune challenge. We thus considered other immune challenges as valid for matching purposes if they occurred the same way, which does not include the relationship between sexual signals and quick immune evaluations (e.g. PHA responses, blood tests).

Guindre-Parker & Love (2014)

<u>Originally reported</u>

"We surveyed published studies on condition-dependent melanin plumage. We characterised species as either achromatic (primarily black, brown, grey or white) or chromatic (also possessing a carotenoid-based red, orange or yellow ornament) to investigate whether the presence of a carotenoid-based signal can influence the condition-dependence of a melanin signal. We excluded species with iridescent plumage as it is primarily structurally based, as well as because it is difficult to determine how birds perceive iridescence without more complex visual modelling."

<u>Issues</u>

1) Unclear how authors verified which sexual signals were melanin-based.

2) The manuscript does not specify proxies included (i.e. what condition-dependent entails), so only proxies present in the dataset were considered to be valid (i.e. hidden criterion for matching purposes).

Hegyi et al. (2015)

Originally reported

“We therefore used correlative studies and took into account alternative explanations arising from the non-experimental situation when interpreting our results. [...] Our first goal was to assess the overall evidence for a correlation between ornamentation and the feeding rate of the ornament bearer or its partner. […] Effect sizes were converted so that the positive sign indicated a positive relationship of feeding rate with plumage colour expression. [...] Due to their negligible number, we had to remove results on incubation feeding of females by males (five results from three populations of three species).

Our search protocol also encountered nine results from four populations of one species (the barn swallow) that concerned feather length as an ornament, which were removed from the dataset. Of the remaining 75 effect sizes, we removed results on white patches due to their prohibitively small number (n=2 from a single species) and also retained for each population of a given species only one measure of the same colour type (n=1 point removed; we chose the point with the more complete information). For each ornament analysed here, the proximate origin of colour variation (carotenoid, pheomelanin, eumelanin, structural) was clear from the species-specific literature. [...] Plumage colour had been measured in different ways depending on the colour category and study approach, and the following changes were regarded as increased colour expression: for carotenoid colour, increased saturation, chroma or patch size; for pheomelanin colour, increased saturation, increased chroma, increased patch size or reduced brightness; for eumelanin colour, increased patch size (no other measures were used in any relevant study); and for structural colour, more UV-shifted hue or increased saturation.”

Issues

1) The manuscript mentioned that only non-experimental approaches were valid but the dataset included cases of experimental approaches. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Hernandez et al. (2021)

Originally reported

"[…] we evaluated the evidence of an association of colourful female ornaments with the condition, reproductive performance and male preferences. […] Three moderators were considered in the model evaluating the association between female colour and condition: Condition proxy (residual body mass, immune response -humoral or cellular -, and parasite load), ornament type (feathers or integuments), and coloration type (carotenoid-dependent, melanin-based, or structural). In the model evaluating the association between female colour and reproductive performance, we used the same ornament type and coloration type moderator variables but replaced the condition proxy with a reproductive performance proxy (laying date, clutch size, or fledging success). When clutch size and fledging success were assessed in the same study, only fledging success was considered. In the analyses exploring the relationship between female colour and male mating preferences, only four effect sizes were available for melanin-based and structural colours (one and three, respectively), so we fit this model including effect sizes only from studies evaluating carotenoid-dependent colorations (n = 11). The ornament type (feathers or integuments) was included as a moderator.

The intensity of coloration was estimated using the number and/or size of coloured structures/patches, spectrophotometry, digital image analysis (RGB or LBA), and visual rank scales (colour charts and visual rank score)."

Issues

1) The manuscript mentions immune responses as valid measures for inclusion, but does not discuss which ones exactly. We only included those which we found matches in the dataset (i.e. hidden rule), which are phytohemagglutinin response and hematocrit.

2) The manuscript mentions that only fledging success was extracted from studies that assessed both clutch size and fledging success. However, we detected cases in which both of these measures were present in the reported dataset from the same study (e.g. [176,241]). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

3) The manuscript mentions that only laying date, clutch size, and fledgling success were valid measurements of reproductive success. However, the dataset contained a case in which breeding success (binary variable) was mislabelled as clutch size. The dataset also contained cases with hatching date (e.g. [241]) and date of first nest (e.g. [206]). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

4) The manuscript mentions that valid condition measures are "residual body mass, immune response -humoral or cellular -, and parasite load", but dataset also includes cases in which body mass is labelled as a condition measure (e.g. [164,167,178]). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Jennions et al. (2001)

Originally reported

“We calculated the effect size as the Pearson product-moment correlation coefficient (r) between trait expression and a measure of survival rate. […] The estimate of survival was based on a comparison between living and dead individuals or an estimate of adult life span/"days alive".

Secondary sexual traits were then scored as being morphological or behavioural characters. We made these comparisons because secondary sexual characters may display a higher degree of condition dependence than ordinary morphological traits. If so, they should be associated with more positive effect sizes."

Issues

1) The manuscript does not contain information on which sexual signals were considered valid for inclusion. We thus considered all sexual signals as valid for matching purposes.

Koch et al. (2016)

Originally reported

"We included only studies (1) reporting the level of carotenoid supplementation as well as the food source provided; (2) including data on both carotenoid-supplemented and control groups of individuals; (3) reporting the values of plasma carotenoid levels and/or coloration; (4) not repeating measures on the same group of birds that were reported in a study already incorporated into the meta-analysis (a potential source of pseudoreplication); (5) testing adult male birds rather than nestlings (in which both carotenoid physiology and ornamental function differ greatly from sexually reproducing adult birds, and the quantity of carotenoids acquired from egg yolk or parental provisioning is often unknown); and (6) supplementing with only the carotenoids lutein and/or zeaxanthin, the most prevalent carotenoid pigments in the avian diet. With the exception of one study supplementing with only lutein, all studies included in our meta-analysis supplemented primarily with lutein and trace amounts of zeaxanthin (e.g., 20:1 ratio of lutein:zeaxanthin). [...] Because 16 of 19 studies investigated songbird species (order Passeriformes), we excluded one study of red junglefowl (*Gallus gallus*), one study of mallards (*Anas platyrhynchos*), and one study of kestrels (*Falco tinnunculus*) to capture the majority of available data while avoiding comparing data from phylogenetically distant taxa with different physiologies. We also excluded one study on society finches (*Lonchura striata domestica*) because this species lacks carotenoid-based ornamentation and so is not subject to the potential costs of allocating carotenoids as colourants.

In addition, because the colour of feathers is determined only during moult when carotenoids are actively deposited in growing feathers, we extracted plumage colour effect sizes only from studies of moulting individuals; we calculated effect sizes from non moulting birds with plumage ornaments only for the relationship between carotenoid intake and plasma carotenoid concentration. The colour of a soft part, such as the bill, can change rapidly during any season, so we could extract both coloration and plasma carotenoid level effect sizes from studies of these ornaments, regardless of moult status. The means of assessing ornamental coloration is important to consider in our analysis because colour is generally quantified along one or more of three main axes: hue, or the shade of the colour (e.g., red, orange, yellow); chroma, or the intensity of the colour (also called saturation); and brightness, or the lightness/ darkness of the colour. In addition, principal component analysis can be used to create a composite metric directly from the reflectance spectrum of a colour. Each of these axes of colour tends to relate to different properties of the coloured ornament itself. For example, chroma may be a good generalisation of pigment density, while hue may be more representative of the proportion of red to yellow pigments in a carotenoid-coloured ornament."

Issues

No issues detected.

Mautz et al. (2013)

Originally reported

“We were interested in four assays of ‘ejaculate quality’: sperm number, sperm swimming speed, sperm size, and sperm viability. We did not include measures of non-sperm characteristics of ejaculates such as seminal chemicals, even though these might affect female fertility and/or male competitiveness under sperm competition, because these relationships are far less clear than those for the four ejaculate traits we do consider. [...] Traits included in each subcategory include: (A) quantity: spermatocrit, sperm count/number, sperm density; (B) size: head length, midpiece length, flagellum length, total length, relative midpiece length, head area, midpiece area; (C) speed: average velocity, average path velocity, curvilinear velocity, linear velocity, straight-line velocity; (D) viability: absolute live sperm, longevity, per cent live, percent motile, percent normal, viability.

[...] papers had to address our main study question: is there a relationship between the expression of a male sexually secondary characteristic (SSC) and an ejaculate characteristic? Authors of the focal papers defined a range of traits as SSCs, including horn size, plumage colouration, song rates, courtship rates, social status, condition and body size. As we were searching for studies related to sexual selection, we relied on the authors of the original paper to define a trait as a SSC (and, by extension, a trait that has been, or still is, under sexual selection). We were specifically interested in sexually dimorphic traits and only included traits that are, generally speaking, not expressed by females (e.g. male courtship colours or advertisement calls) or differ qualitatively between the sexes (e.g. major differences in horn shape). We excluded traits such as social dominance, body condition and, most importantly, body size that are not usually treated as SSCs. [...] We were interested in SSCs rather than attractiveness per se so we exclude studies that only report on composite measures of attractiveness (e.g. ‘mating latency’).”

Issues

No issues detected.

McLean et al. (2012)

Originally reported

"We carried out an extensive literature search (completed in April, 2010) to find studies that documented the inter-sexual relationships among the three traits in anuran species: (1) the correlation between male call frequency and male body size (hereafter, the frequency–size relationship), (2) female preference for male call frequency (the preference–frequency relationship) and (3) female preference for male body size (the size–preference relationship). [...] For the relationships that involved female preferences, we included studies that quantified the relationship between male mating success and signal or quality traits, assuming differential male mating success was generally the result of female choice."

Issues

No issues detected.

Meunier et al. (2011)

Originally reported

"For each study, sign and magnitude of the correlation between melanin-based coloration and laying date, clutch size, brood size and survival were given by the parameter 'effect size' r calculated following standard methodology. We defined a positive effect size when individuals with larger or darker melanin-based colour traits had lower laying date, had larger clutch size, larger brood size and higher survival rate than individuals with smaller or paler melanin-based colour traits. [...] We restricted our meta-analyses to eumelanin-based coloration (i.e. black and grey coloration) because few studies have yet been published on pheomelanin-based coloration (i.e. reddish-brown coloration).”

Issues

No issues detected.

Moller & Alatalo (1999)

Originally reported

“The aims of the present study were to quantify the viability effects of sexual selection. This was carried out based on a literature survey of studies of good-genes sexual selection. Although good-genes effects may be expressed as enhanced growth, fecundity or survival, we have concentrated our efforts on reviewing the literature on survivorship effects because most studies have addressed this major fitness component, and because life-time reproductive success in a diverse array of organisms depends more on longevity than on any other life-history trait. [...] The variables of interest were classified in the following ways: (i) whether the target of selection had been identified based on observations or experiments, or whether that was not the case; (ii) the magnitude of the viability effect, calculated as the correlation coefficient between a secondary sexual character and viability of the offspring; (iii) the female mate preference for the male trait estimated from observational or experimental studies, expressed as the correlation coefficient between the male character and male mating success [...].”

Issues

1) The manuscript does not contain information on which sexual signals were considered valid for inclusion. We thus considered all sexual signals as valid for matching purposes.

Moller & Jennions (2001)

Originally reported

“We determined the relationship between the expression of male secondary sexual characters (or other characters associated with male mating success that appear to influence female choice decisions) and four components of direct fitness for females. [...] First, fertility was determined as either (1) the proportion of eggs fertilised among females mated to a given male; (2) the probability that copulation led to offspring production; or (3) the proportion of eggs that hatched in species without paternal care. Since multiple mating is common among females of many species and sperm storage may occur, we can be sure that this effect is larger than estimated here. Second, we determined fecundity as (1) clutch size; (2) the number of eggs laid over a specified time interval; or (3) litter size in live-bearing species. Third, we determined male investment in parental care in birds as (1) the proportion of feeding visits to offspring relative to the total number of feeds by both male and female, or (2) the absolute feeding rate of the male. Where possible we used the latter measure because differential allocation by females may lead to an underestimate of the absolute amount a male invests in offspring. Four, we also included measures of hatching success for fish, amphibians and insects where there is paternal care but males do not feed offspring. In all these species it is clear that male parental behaviour is a major factor in the hatching success (e.g. due to fanning of eggs or attacking predators). We excluded data on fledgling production or the number of young reared to independence by birds and mammals because we did not feel we could adequately cover the huge literature that this entails. However, we are unaware of any study relating the expression of secondary sexual characters to the quality of male parental care and subsequently to offspring viability (recruitment into the following reproducing population). In addition, variation in territory quality is likely to play a major role in offspring survival because it will influence the rate at which offspring are fed. Thus, we would need to partition reproductive success into effects related to the expression of male secondary sexual characters and effects due to territory quality per se. We are unaware of any studies that have done so.

We treated eggs in the nest as a sexually selected character in fish. Exclusive male parental care in insects (and fish) appears to be a sexually rather than naturally selected trait, because females should use the number of clutches or eggs a male cares for as a reliable signal of paternal intent and quality."

Issues

1) The manuscript does not contain information on which sexual signals were considered valid for inclusion. We thus considered all sexual signals as valid for matching purposes.

2) The manuscript mentions that "excluded data on fledgling production or the number of young reared to independence by birds and mammals" but we detected cases in the dataset with number of fledglings as a proxy in bird species (e.g. [151]). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

3) Despite the criteria listed in the manuscript, occurrence of a second clutch as a reproductive success measurement was detected in a few cases of the dataset reported. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Moore et al. (2016)

Originally reported

"We only included those studies in which the following criteria were met: 1) subjects were adults; 2) subject sex was specified; 3) physiological indices of stress were measured […] Four categories of stress measurement were reported: baseline GCs, peak or total GCs produced in response to a stressor, experimental elevation of GCs, and long-term stress. Baseline GCs were typically measured within 3–5 min of capture. Experimental elevation of GCs up to 4 times above baseline was achieved via subcutaneous implants containing GCs. Long-term stress was assessed in 3 ways: GCs deposited in feathers, faeces or hair; the ratio of heterophils to lymphocytes (a white blood cell count that correlates with baseline GCs); and the expression of heat shock proteins (highly conserved proteins that are elevated under stress). Both heterophil-to-lymphocyte ratio and heat shock proteins are widely used as proxies of recent and long-term stress in the ecological literature.

The effect sizes we obtained considered a wide range of secondary sexual traits, which we sorted into 4 categories: coloration, vocalisation, morphological traits, and opposite-sex preferences. The coloration category included examples in birds, mammals, and reptiles. The amount of coloration was measured in several different ways, including brightness, hue, saturation, proportion of structure (e.g., eye ring) that is pigmented, ultraviolet reflectance, and colour reflectance. The vocalisation category included singing in birds and calling in amphibians and a mammal species (rock hyrax *Procavia capensis*). The parameters measured varied according to the nature of vocalisation in each species and included song rate, complexity, and repertoire size in birds; the latency to call, call duration, call rate, and vocal effort in amphibians, and whether calling/singing was observed or not (rock hyrax, amphibians). Effect sizes included in the morphological trait category all considered bird species and assessed the size of secondary sexual characters, such as comb or tail length. We also included in this category effect sizes considering the size of a coloured structure (but not the coloration itself), such as epaulet size in the red-winged black- bird (*Agelaius phoeniceus*). Finally, while not a secondary sexual trait per se, opposite-sex preference was included as an indirect measure of the level of sexual signalling, with the assumption being that attractiveness to the opposite sex is a function of investment in secondary sexual traits. We rely on author judgments regarding whether each trait is a secondary sexual trait or not.”

Issues

No issues detected.

Nakagawa et al. (2007)

Originally reported

“We conducted an extensive search of the literature on the relationship between life-history traits and bib size in house sparrows. […] Our criteria resulted in 6 life-history traits that could be investigated: fighting ability, parental ability, age, body condition (i.e., standardised weight), cuckoldry (the rate or absence/presence of lost paternity due to extra pair paternity [EPP]), and reproductive success. In the analysis of parental behaviour, we used studies investigating both incubation and feeding behaviour. A recent study showed incubation time to be a significant predictor of subsequent nestling provisioning in a population, so that treating incubation and provisioning together may be justifiable.”

Issues

1) The manuscript mentions that reproductive success is a proxy without any details but their dataset only contains the number of fledglings (even though other measures were available, e.g. number of eggs, number of hatchlings). Thus, this was considered a hidden criterion for matching purposes.

Nolazco et al. (2022)

Originally reported

“Condition parameters fell into six categories: (1) body condition: mainly measurements of body mass adjusted by structural body size and others associated with the physical condition of individuals; (2) body size: structural size (measurements of tarsus, wing, beak, keel, and tail alone or in combination) and mass; (3) immunity: indicators of constitutive immunity, immune challenges and responses; (4) stress: indicators of baseline physiological stress, stress challenges, and capacity to cope with oxidative stress; (5) environment: climatic conditions and resources; and (6) parasites: incidence and abundance of parasites. [...] Fitness parameters included not only estimates of reproductive success, survival and offspring quality, but also parental quality (because parental investment has been hypothesised to vary as a function of ornamentation) and timing of breeding. Hence these factors were classified into five categories: (1) reproductive success: mating success and offspring production; (2) offspring quality or condition: measurements of egg quality, offspring body condition, immunity, parasites, and other indicators of physical condition; (3) parental quality: provisioning during incubation, and offspring feeding and defence; (4) timing of breeding: measured directly or as arrival time to breeding grounds; and (5) survival.

We applied a phylogenetically controlled bivariate meta-analytic approach to quantify the strength and direction of associations between ornaments and condition or fitness, in bird species in which both sexes were ornamented. Only morphological ornaments that were visually recognisable and identified as such by the authors of the original studies were included, excluding traits that did not appear decorative such as body size and weapons. In all cases, ornaments were similar in structure and location between sexes. We note that in most cases ornamental function of these traits is assumed and experimental evidence is not available. Thus, we broadly defined ornaments as any phenotypic traits that look like decorations rather than having an apparent naturally selected function. [...] We excluded publications if only male traits were investigated (i.e., no conspecific female traits were studied in this or another publication). For any female-only studies, we used the cited reference list to identify the corresponding information on conspecific males, adding 25 additional publications. Our criteria for including these studies was that the data originated from the same populations. [....] We classified ornamental traits into six categories: (1) carotenoid-based colouration: yellow, orange, or red coloured ornaments; (2) melanin-based colouration: black, grey (eumelanin), or brown (pheomelanin); (3) structural-based colouration: iridescent and non-iridescent; (4) unpigmented: white patches; (5) morphological: morphology of ornamental appendages (e.g., comb, wattle, tail, plumes), and (6) others: to cases in which the operational variables were a combination of two or more ornament categories or rare pigments (e.g., spheniscin in penguins)."

Issues

No issues detected.

Parker (2013)

Originally reported

"I attempted to locate all papers published in English which assess the function or characteristics of plumage colour in blue tits. [...] No clear expectations about seasonal changes in colour emerge from sexual selection hypotheses, and so I did not assess tests for these patterns. […] there is no clear prediction about the hue (wavelength of peak reflectance) of white plumage because it is characterised by relatively uniform reflectance across the spectrum visible to birds. Carotenoid-pigmented avian plumage requires dietary carotenoids regardless of any role in sexual selection and so I did not consider the effect of manipulation of dietary carotenoid intake on the colour of the carotenoid-pigmented yellow breast to be a test of a sexual selection hypothesis. I also excluded the one observational study that used a qualitative colour score rather than a quantitative colour measurement. [...] Age: Differences in plumage colour between second-year (yearling) and after-second-year individuals; Aggression directed:

Aggression predicted by the plumage colour of the aggressor; Aggression received: Aggression predicted by the plumage colour of the victim, including the relationship between plumage colour manipulation and resulting dominance rank; Aggression dominance: Relationship between plumage colour and dominance rank; Mate choice – assortative: Correlations between plumage colour of one member of a mated pair and plumage colour or other morphological trait of the other member of mated pair; Mate choice – differential investment: Parental plumage colour related to the mate's investment in reproduction; Mate choice – EPP: Male plumage colour and rates of extra-pair paternity (EPP) determined by molecular paternity analysis; Mate choice – WPP: Male plumage colour and rates of within-pair paternity (WPP) determined by molecular paternity analysis; Mate choice – trials: Captive mate-choice trials; Quality: Plumage colour and measures or manipulations of individual quality or measures of individual's offspring quality; Sex: Differences in plumage colour between males and females; Sex ratio: Male plumage colour and sex ratio of offspring."

Issues

No issues detected.

Parker & Ligon (2003)

Originally reported

"We were interested in testing two basic hypotheses: (a) female mating decisions correlate with male comb morphology, and (b) female mating decisions correlate with male feather morphology."

Issues

1) The dataset was unclear regarding sexual signals (e.g. whether colour was hue, chroma, brightness, etc).

Parker et al. (2006)

Originally reported

"We located each published study comparing blue tit song with potential indices of individual quality or condition."

Issues

No issues detected.

Prokop et al. (2012)

Originally reported

"We collected studies reporting on either (1) genetic correlations between male sexual traits and other fitness-related characters, estimated using quantitative genetic methods (animal model, full-sibling/half-sibling designs), (2) correlations between sire sexual traits and offspring fitness-related traits, or (3) heritability of male sexual characters. [...] We also included studies where measures of male mating success (e.g., comparing males that did and did not achieve copulations in mating trials) were used instead of specific sexual traits, provided that (1) the success could be attributed to female choice rather than/apart from male–male competition and (2) different individual females were used to determine mating success of any given male and to produce his progeny scored for fitness traits. [...] Included only when paternity had been experimentally controlled or genetically confirmed—or the frequency of extra-pair offspring was known to be <15% in the population studied (<20% if the sample size was at least 200), as such levels of extra-pair paternity should not bias the estimates of genetic parameters. [...]

In (1) and (2), we included data from species with multiple sexually selected traits, where pairs of such traits were correlated with each other. In all cases, we only took into account male sexual traits known or supposed to be targets of female choice, therefore excluding data on traits used only for intrasexual competition for mates. We also included studies where measures of male mating success (e.g., comparing males that did and did not achieve copulations in mating trials) were used instead of specific sexual traits, provided that (1) the success could be attributed to female choice rather than/apart from male– male competition and (2) different individual females were used to determine mating success of any given male and to produce his progeny scored for fitness traits. [...] We classified sexually selected traits as display (purely ornamental) and competitive (size/dominance related); a third category, "other," consisted of traits that could not be assigned to either of the two (such as nuptial gift, mating success that could not be specifically attributed to either display or dominance, or a trait increasing postcopulatory success)."

Issues

No issues detected.

Robinson & Creanza (2019)

Originally reported

"We obtained field studies that examined the link between individual song elaboration (number of songs or syllables) and reproductive success (reproductive output or mating success). From the 48 field studies that remained, we compiled 134 relevant measurements. These commonly used measurements of reproductive success were categorised as follows: Number of females:

how many social mates a male attracts, where males who attract more females are assumed to be more successful. Latency to pairing date or laying date: these two measures are traditionally used as a measure of reproductive success, because attractive males should pair first, and those who produce offspring sooner have a better chance of parenting a larger brood or more than one brood of offspring in a single breeding season. Chicks born earlier in the season also tend to be more viable. Extrapair paternity: this is often considered a metric of reproductive success because males that sire offspring in extra pair matings are assumed to be more attractive to females. However, it has been suggested to be an unreliable metric. Clutch size or number of offspring/recruits: these three measures are affected by both male and female genetic quality; however, it has been shown that females exposed to more elaborate songs can respond by producing larger clutches, so male song quality can also potentially affect this metric. The number of offspring or number of recruits (offspring that return to the parental territory) is related to the genetic fitness of males and females, but also to parental investment.

In the studies that remained, individual song elaboration was measured by either song repertoire size (unique number of songs per individual) or syllable repertoire size (unique number of syllables per individual). We included studies that measured the association between reproductive success and either of these song elaboration metrics, because syllable repertoire size and song repertoire size are correlated between species and are likely also correlated within species."

Issues

1) We detected one data point mislabeled as number of recruits when it should be residual recruitment (i.e. controlled for number of fledglings, a measure of viability rather than reproductive success). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Romano et al. (2017a)

<u>Originally reported</u>

"In our data set we only included the statistical relationships where plumage ornaments were hypothesised to determine fitness-related traits, independently of how the test was designed. Thus, we included the results of analyses where an ornament was included as a predictor of a fitness trait, as well as analyses testing for a difference in ornament expression between groups of individuals with different fitness (e.g. tests comparing tail length of mated versus unmated individuals). Importantly, we did not include tests of condition-dependence of the expression of ornaments (e.g. tests of variation in ornament expression according to parasite load, physiological parameters or environmental conditions) nor tests of the potential costs of ornaments, because these were not the focus of the study. [...] Briefly, data on reproduction were analysed by comparing relationships between each plumage ornament and both male and female reproductive output by separately considering the following fitness proxies, which refer to different phases of the breeding cycle: (i) mating success, which comprised the probability of obtaining a social mate, and the time elapsed between arrival date to the breeding site and reproduction; (ii) mating date; (iii) success in paternity, as gauged by gaining extra-pair and within-pair offspring, as well as by successfully engaging in extra-pair copulations; (iv) laying date; (v) breeding success, including clutch size, brood size, and fledging success for any breeding attempt; and (vi) overall reproductive success, considering the number of broods, all eggs produced, and total number of offspring sired (but not when total number of eggs/offspring was weighted by the number of broods) during the entire breeding season. We note that data on clutch and brood size were pooled because in the barn swallow both the number of unhatched eggs and mortality rate during the nestling period are normally very low. In practice, the number of eggs laid and the number of nestlings fledged are highly correlated. Data on parental care were divided into two categories, corresponding to different phases of the breeding period: (i) incubation period, including the duration of incubation and the (absolute or relative) time spent by females or males in incubating eggs; and (ii) care provisioning of nestlings, including feeding rate, number of prey brought to the nest, and duration of the nestling period. Importantly, we distinguished between parental care provided by the mother and the social father of the nestlings. We then identified two categories of offspring quality: (i) offspring size, including skeletal size (e.g. tarsus length) and body mass measurements of nestlings; and (ii) offspring physiology, accounting for immune function and other physiological variables. Finally, we also recorded effect sizes concerning the associations between arrival date or survival/mortality and plumage ornaments. Importantly, data regarding laying date, incubation, breeding success, care provisioning and offspring quality were categorised according to the breeding attempt to which they referred, because barn swallows often lay more than one clutch per breeding season, and the intensity of sexual selection may vary during the breeding season. We thus considered separately the results concerning first or second broods (no data were available for subsequent broods, as very few females lay more than two clutches in a breeding season). Data on paternity may refer to the first brood only or to both broods pooled. Because the fitness proxies could be ordered chronologically, we could test whether the intensity of sexual selection varied among different phases of the breeding cycle. We note that ‘breeding success’, indicating clutch and brood size, was placed before ‘care provisioning’ because it mainly reflects parental decisions on initial clutch size rather than subsequent adjustments of brood size (see above). Analysis of variation in the effect size among different phases of the breeding cycle was limited to first broods because of the small number of effect sizes (and breeding stages) concerning second broods. In addition, because only high-quality breeding individuals (i.e. the more ornamented ones) usually lay a second clutch, during second broods smaller inter-individual variability in ornament expression compared to the first broods is expected. This difference in ornament variation between first and second broods may therefore reduce mean effect during the breeding season.

In our analyses we considered the following sexually dimorphic characters: tail length, tail asymmetry, size of white spots on tail, ventral plumage colour, throat patch colour and size."

Issues

No issues detected.

Romano et al. (2017b)

Originally reported

"We included information about both annual survival (e.g. probability to survive from one breeding season to the next) and survival during the breeding season (e.g. survival from predators and severe weather). We emphasise that in our model species, both breeding dispersal and emigration from one breeding site to another are extremely rare. Considering that all the studies included in the analyses have been performed during the breeding season, the effects of dispersal and emigration on individual survival are therefore virtually absent.

We considered the published studies where survival was tested in relation to the expression of six sexually dimorphic characters which have been suggested to be under sexual selection in at least one of the barn swallow subspecies: tail length, tail asymmetry, size of white spots on the tail, ventral plumage colour, throat plumage colour, and throat colour patch size."

Issues

No issues detected.

Sanchez-Tojar et al. (2018)

Originally reported

“We only included articles in which dominance was directly inferred from agonistic dyadic interactions over resources such as food, water, sand baths or roosting sites.

When the presence of multiple estimates was due to the use of different methods to estimate bib size and dominance rank on the same data, we chose a single estimate per group of birds or study based on the order of preference [...]” [exposed/visible bib > hidden bib > total bib].

Issues

No issues detected.

Santos et al. (2011)

Originally reported

“The study must have: (1) investigated the relationship between dominance and plumage traits in one sex, or statistically accounted for sex effects on the dominance interactions (such sex effects are well known, and must be accounted for so that the relationship between dominance and plumage is not confounded); (2) reported whether the plumage characteristic of interest was manipulated or not; (3) reported whether interactions observed to assess dominance were territorial or over food resources; and (4) fully reported test statistics and their associated degrees of freedom (or sample size) for both significant and nonsignificant results. As a result of applying our first inclusion criterion, our data consist only of the relationship between dominance among males and male plumage (i.e. no data were available on the relationship between dominance and plumage among females).”

Issues

No issues detected.

Simons & Verhulst (2011)

Originally reported

"To test whether the correlation between attractiveness and bill color can be attributed to a correlation between bill color and song rate, our approach was to quantify the association association between bill color and song rate using meta-analysis on a different set of studies and compare the strength of this correlation with the correlation between the color of a male's bill and his attractiveness.

The statistical approach between studies differed, with some reporting the preference for the reddest male and others reporting the relationship between the difference in redness and the resulting female preference. The second approach includes both the effect of the difference between males in redness together with the overall preference for the reddest male. We recommend reporting both in future research to ease comparison between studies. For the purpose of this review, we included both approaches because the rejection of either approach would have resulted in a substantial loss of studies. We preferred the statistic of the preference for the reddest males if both approaches were available."

Issues

1) The manuscript specified measures of body condition as size, size-corrected mass, or growth rate, but the dataset also included others like subcutaneous fat score. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Simons et al. (2012)

Originally reported

“We summarised five phenotypic relationships: circulating carotenoid levels with trait redness, immune function and oxidative stress state; and trait redness with immune function and oxidative stress state. The relationships with trait redness represent signalling value, i.e. the information that can be obtained by a choosing individual regarding the physiological state of the signaler. The relationships with carotenoid levels represent the hypothesised mechanisms maintaining signal honesty. [...] Exclusion criteria: [...] iii) An immune challenge or oxidative stress challenge was given after which carotenoid levels or sexual coloration were assessed. Our focus here is whether carotenoid levels or carotenoid-dependent coloration predict oxidative stress parameters or immune response. The question of whether challenges reduce carotenoid levels or redness of sexual coloration is relevant, and this mechanism may in part or fully underlie between-individual variation in sexual coloration. However, the effects of experimentally induced immune or oxidative stress cannot be directly scaled to natural variation or direct manipulation of carotenoid levels and may involve different trade-offs and hence we excluded such studies. iv) When carotenoid supplementation was applied experimentally, but data on natural variation in circulating carotenoid levels or coloration were also available, we used the latter because this is the variation that a choosing potential mate is confronted with. [...] We included [...] whether the effect size was subject to experimental variation, caused by treatments other than carotenoid supplementation, which potentially increased variation in the traits of interest. To avoid such effects we selected pre-experimental (including carotenoid supplementation studies) values or results of analyses of the control group only, when possible. [...] In our analyses we considered the measures of the immune system of which we found four or more independent studies. These measures were as follows: PHA response, antibody production against experimentally induced antigens, parasite load, and white blood cell counts.”

Issues

1) The manuscript reports that only circulating carotenoids in plasma were included as a measure of body condition and the dataset reports the same (e.g. "carotenoid level"). However, we noticed that studies with distinct carotenoid diets appear to have been equated as carotenoid level. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

2) The manuscript focuses on redness as a sexual signal and mentions chroma and hue as valid measurements. However, we noticed that the extent of a colourful trait was also used in some occasions and that traits could be of other colours than red (e.g. yellow). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

Soma & Garamszegi (2011)

Originally reported

“We used the following criteria for inclusion of studies: 1) results on the relationships between song complexity and reproductive success were based on observational data in wild birds without experimental manipulations; 2) song complexity was measured by either song or syllable repertoire size (i.e., the number of unique song or syllable types in the full repertoire or in a given amount of vocal sample), versatility, or principal component score of multiple song variables that reflect song complexity; 3) the measure of reproductive success was based on either the number of females mated (including both social and extrapair mating), the number of offspring sired (including the number of eggs, fledglings, and recruits), or the timing of the onset of reproduction.”

Issues

No issues detected.

Street et al. (2016)

Originally reported

"To fulfil the inclusion criteria, articles had to contain either (1) data on changes in swelling size across days relative to an independent estimate of the day of ovulation based on ovarian hormone levels (from blood, urine or faecal samples) or laparoscopy, or (2) data on swelling size and any measures of individual female quality (e.g. body condition or social rank). [...] Only those measures that were reported in at least four studies were analysed, i.e. age, social rank and body condition.

[…] data from species with small swellings (defined as swell- ings that involve the vulval and clitoral areas only) were excluded (i.e. *Hylobates lar*; *Gorilla gorilla*), as were data from species with subcaudal swellings (i.e. *Macaca fas*). The relationship between swelling coloration and ovulation was not investigated, as few studies with suitable data were available."

Issues

No issues detected.

Thornhill & Moller (1998)

Originally reported

"Because we investigated the relative effect size for symmetry versus character size for the same trait, the only studies included from Moller and Thornhill (1998) were those that had one or more estimates that allowed calculation of an effect size for the relationship between asymmetry and character size, and mating success or a mate preference, respectively."

Issues

No issues detected.

Weaver et al. (2018)

<u>Originally reported</u>

“We quantitatively synthesised published results on the relationships between individual quality and plumage coloration of passerines produced via dietary versus converted carotenoids. […] studies must have investigated at least one of the following proxies of individual quality: (1) nutritional condition, (2) immune function or oxidative capacity, (3) parasite resistance, or (4) reproductive or parental quality.

[...] we focused our study on a single trait, feather coloration, to avoid confounding biological factors such as blood flow, carotenoid esterification, and differences in requisite enzymes that are relevant to bare-part coloration, but not feather coloration. [...] Because we were interested in the signal content of carotenoid-based plumage, we focused on studies that quantified feather colour using standardised colour metrics of natural (i.e., unsupplemented) adult bird colour levels. Therefore, we excluded studies from our meta-analysis for any of the following reasons: only coloration of non-feathered structures was measured (e.g., wattles, legs, beaks); a non-passerine species was studied; only plasma concentrations of carotenoids were measured; or only nestling or juvenile coloration was studied. We did not include measures of feather brightness as it is sensitive to factors unrelated to pigmentation. [...] Common metrics used to quantify this variation in feather reflectance include comparisons to standard colour charts (e.g., Munsel), calculations of hue, chroma and brightness or composite metrics such as principal components (PCA) from spectrophotometer data or digital photographs. Hue describes the unique spectral colour (e.g., “red”, “orange”, “yellow”) and chroma describes the saturation or spectral purity of the colour display relative to total reflectance across the visible range of the electromagnetic spectrum. [...] We did not include measures of brightness in our analyses because it is strongly influenced by the physical structure of the feather which may be altered by abrasion and wear and is difficult to interpret for carotenoid content."

Issues

1) The manuscript states that only colour metrics are valid sexual signals but dataset reveals that also included size of colourful plumage (e.g. patch size). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

2) The manuscript states that only adult sexual signals were used but the dataset includes data from juveniles from many empirical studies. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

3) The dataset includes offspring size but states it as "offspring quality" even though this is omitted in the manuscript. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

White (2020)

Originally reported

"I included all experimental and observational studies that quantified the relationship between intersexual structural colour signal expression (via the measurement of hue, saturation or brightness, or a composite thereof) and any one of age, body condition (size, size-corrected mass or growth rate), immune function (oxidative damage, PHA response, circulating CORT or testosterone) or parasite resistance as a measure of individual quality."

Issues

1) The manuscript states that only structurally-based sexual signals were included but some carotenoid-based ornaments were included in the dataset. Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

2) Data on distinct carotenoid diets were apparently used as condition proxy, even though the manuscript lists only other proxies for "condition" ("size, size-corrected mass or growth rate"). Nonetheless, we followed the original rule given in the manuscript for matching purposes for other studies.

3) The manuscript mentions that the condition proxy could be "size, size-corrected mass or growth rate", but only body mass was used as a body size measurement, which we interpreted as a hidden criterion.

4) Unclear how colours of sexual signals were assessed to be structurally-based. Description of sexual signals in the dataset only specify colour measurement (e.g. hue) but not body part.

Yasukawa et al. (2010)

Originally reported

“We performed a meta-analysis to determine whether epaulet expression affects male-male competition, female choice, or reproductive success. This analysis provides a quantitative way to interpret the wide range of results from a variety of observational and experimental studies of Red-winged Blackbird epaulets as sexually selected adaptations.”

Issues

1) Specific included measurements for aggression are unclear in the manuscript and dataset given by authors contains only vague descriptions of aggression (number of hits, close approaches, approaches, dominance rank) but more were available in empirical studies from

where data were extracted. We interpreted that only those reported were considered valid, and thus as a hidden criterion for matching purposes.

Supplementary information S5. Examples of matching attempts

The easiest attempts to find the correct matches between datasets occurred in cases in which a given empirical study (source for data extraction) contained only one data point with the same description (i.e. same sexual signal and proxy) in the originally reported dataset as well as in the re-extracted dataset. For instance, Simons et al. [47] reported a data point from Birkhead et al. [257] on the relationship between song rate and bill redness in zebra finches. In our re-extracted dataset, only one data point existed for this empirical study on this same relationship, so there was no doubt that it was a match to the data point originally reported. Unfortunately, most cases were not as simple. For example, Santos et al. [46] reported a single data point from Korsten et al. [119] on the relationship between plumage colour and territory defence in Eurasian blue tits. Yet, our re-extracted dataset contained 38 data points for the relationship between plumage colour and aggression or social dominance. Luckily, only two of these mentioned territorial defence, with one of them with the same sample size as the one reported. If the proxy in the dataset of Santos et al. [46] was slightly more vague (aggression instead of territory defence) or if the sample size reported was different from the any of our re-extracted data points, then we would have to label all re-extracted data points with the same extraction ID, as if that single reported data point was an average of all possible data points given in the empirical study. Indeed, this is what happened for complicated cases like [213], used by Simons et al. [47], who reported three data points, all on the relationship between bill redness and attractiveness in zebra finches. Our re-extracted dataset contained nine data points with similar description, and no other information was available to clarify this conundrum. We thus matched the three originally reported data points with all nine re-extracted data points. We classified matches one-to-one as exact matches, while other matches (one-to-multiple, multiple-to-one, multiple-to-multiple) as non-exact.

Supplementary information S6. Calculation and conversion of effect sizes

We used Fisher's *Zr* in most of our analyses. Following Hedges et al. [327], we mainly used Pearsons' correlation coefficient (*r*) to calculate *Zr* as:

$$Zr = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

with its sampling variance (se^2) as:

$$se^2 = \frac{1}{n-3}$$

where *n* is the number of individuals tested (i.e. total sample size). If needed, we transformed *Zr* to *r* as:

$$r = \frac{e^{2Zr}-1}{e^{2Zr}+1}$$

Data reported by empirical studies were available in various formats. If *r* itself could not be obtained, we calculated it from other data. For instance, if only the coefficient of determination was available (R^2), we calculated r as:

$$r = \sqrt{R^2}$$

Other types of correlations were directly transformed to *r* following Koricheva et al. [3], such as Spearman's rank correlation (ρ, only needed if $n < 90$, otherwise equates to *r*), as:

$$r = 2\sin\left(\frac{\pi\rho}{6}\right)$$

and Kendall's correlation (τ), as:

$$r = \sin\left(\frac{\pi\tau}{2}\right)$$

Goodman and Kruskal's gamma (γ) is similar to Kendall's correlation, so we used the equation above with γ instead of τ to obtain *r* when necessary.

By contrast, other types of data needed further steps to obtain *r*. For example, following Jacobs et al. [328], *t* from independent samples *t*-tests were transformed to point-biserial correlations (r_{pb}) as:

$$r_{pb} = \frac{t}{\sqrt{t^2 + n - 2}}$$

Similarly, following Koricheva et al. [3], *F*-values were transformed to r_{pb} as:

$$r_{pb} = \frac{\sqrt{F}}{\sqrt{F + n - 2}}$$

Following Jacobs et al. [328], we assume that *r* equals to the biserial correlation coefficient (r_b), which can be calculated from r_{pb} as:

$$r = r_b = \frac{\sqrt{pq}}{f(z_p)} r_{pb}$$

where $\mathrm{p} = \frac{\mathrm{n}_1}{\mathrm{n}}$, $q = \frac{n_2}{n}$, n_1 is the sample size of the first group (e.g. treatment), n_2 is the sample size of the second group (e.g. control), and $f(z_p)$ denotes the density of the standard normal distribution at value z_p, which is the point for which $P(Z > z_p) = p$, with *Z* denoting a random variable following a standard normal distribution. We assumed $p = q = 0.5$ in all instances in which we used this equation.

Other statistics had to be first transformed to *t*, such as (1) regression slopes (β) as:

$$t = \frac{\beta}{se},$$

(2) Chi-Squares (χ^2; but only those with one degree of freedom, i.e. comparisons between two groups) as:

$$t = \sqrt{F} = \sqrt{\chi^2},$$

and (3) *p*-values using the *qt* function (*t* distribution) in R 4.4.0[310].

Differences between groups were also reported in different ways, and most of them could not be directly used to calculate *r*. Instead, we calculated Cohen's *d* from these data before converting them to *r*. Following Borenstein et al. [4], we obtained *r* from Cohen's *d* as:

$$r = \frac{d}{\sqrt{d^2 + a}}$$

where a is:

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2}$$

When only total sample size was given, we assumed n2 = $n_1 = n_2$. Following Cohen [329], descriptive data were first used to calculate Cohen's d as:

$$d = \frac{m_2 - m_1}{s_{pooled}}$$

where m_2 and m_1 are group means (e.g. treatment and control), while s_{pooled} was calculated as:

$$s_{pooled} = \sqrt{\frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_2 + n_1 - 2}}$$

where s^2 is the variance of each group. However, if data from two points in time (e.g. pre- and post-experiment) were given for each of the two groups, following Gurevitch et al. [330], we calculated Cohen's d as:

$$d = \frac{\left(m_{2,post} - m_{2,pre}\right) - \left(m_{1,post} - m_{1,pre}\right)}{s_{pooled*}}$$

while $s_{pooled*}$ was calculated as:

$$s_{pooled*} = \sqrt{\frac{\left(n_{2,post} - 1\right)s_{2,post}^2 + \left(n_{2,pre} - 1\right)s_{2,pre}^2 + \left(n_{1,post} - 1\right)s_{1,pre}^2 + \left(n_{1,pre} - 1\right)s_{1,pre}^2}{n_{2,post} + n_{2,pre} + n_{1,post} + n_{1,pre} - 4}}$$

If only the difference between groups was given instead (as a single value), we calculated Cohen's d as:

$$d = \frac{\overline{m}}{s_{pooled}}$$

where m is the difference between groups. If only results from paired-samples *t*-tests (t_{paired}) were provided, following Dunlap et al. [331], we obtained Cohen's *d* from them as:

$$d = t_{paired}\sqrt{\frac{2(1-r_{12})}{n}}$$

where r_{12} is the correlation coefficient between two groups (set as 0.5 in all instances). If only medians and interquartile range for each group were given (as in boxplots), we used equations 14 and 15, along with table 2, from Wan et al. [332].

Furthermore, data could be reported in the form of frequency of events or proportions. If the former, following Nakagawa et al. [333], we first calculated *r* as:

$$r = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

where A, B, C, and D are frequencies from a two-by-two contingency table. If proportions for two groups were given and they could not be transformed into frequencies then, as in Pollo et al. [316], we calculated Cohen's *d* as:

$$d = \frac{logit(p_2) - logit(p_1)}{\pi}\sqrt{3}$$

where p_2 and p_1 are the proportions reported for each group. Lastly, if only a single proportion was given (e.g. proportion of time spent with one prospective mate over another), as in Pollo et al. [316], we calculated Cohen's *d* as:

$$d = \frac{logit(p) - logit(1-p)}{\pi}\sqrt{3}$$

where *p* is the proportion reported.

We could not compute effect sizes when the only data reported were (1) measures mentioned above without all necessary components to calculate effect sizes (e.g. means without standard errors), (2) genetic correlations, (3) Wald Chi-Square, (4) selection gradients, (5) Mann-Whitney *U*-Tests, and (6) *F*-values from ANOVAs with predictor variables with more than two levels.

Supplementary information S7. Direction of effect sizes

We had to invert the direction of some effect sizes originally reported by two meta-analyses because authors reported these effect sizes in the opposite direction than the one they were analysed: (1) [30] with effect sizes on the relationship between frequency of acoustic signals and body size or mating success, and (2) [42] with effect sizes related to latency to court. Moreover, we obtained the absolute value for all effect sizes from White [53] in which the sexual signal involved colour hue, as the author stated doing so for their analysis.

When re-extracting data from empirical studies, we followed the rationale that individual condition and fitness are reflected by greater sexual signal expression, except for the following sexual signals: asymmetry, inter-pop or intercall interval, latency to signal, song frequency in amphibians (but not in other animals), and other measures (e.g. composite measures, brightness, hue) in specific cases (e.g. lower hue for blue traits). Likewise, the direction rationale we used for various proxies (i.e. how they reflected individual condition or fitness) are shown in Table S1. The direction of data points re-extracted is also reported in detail in our dataset.

Despite how we proceeded when re-extracting data, we adjusted the direction of re-extracted effect sizes to comply with the rationale stated in original meta-analyses. Therefore, we re-inverted re-extracted effect sizes for the following cases because they were in the opposite direction than the one we assigned to them during re-extraction:

1. All effect sizes that could be included by [14], as positive values there represent a positive relationship between sexual signal expression and parasite load.
2. All effect sizes that could be included by [34], as positive values there represent a positive relationship between asymmetry of sexual signals and attractiveness.

3. All effect sizes that could be included by [35], as positive values there represent a positive relationship between sexual signal expression and stress.
4. Effect sizes related to age that could be included by [15], as positive values there represent a negative relationship between sexual signal expression and age.
5. Effect sizes related to parasite load, heterophil-to-lymphocyte ratio, and oxidative damage that could be included by [48], as positive values there represent a positive relationship between sexual signal expression and both parasite load and stress.
6. Effect sizes related to asymmetry of sexual signals that could be included by [51], as positive values there represent a positive relationship between asymmetry of sexual signals and attractiveness.
7. Effect sizes related to parental care that could be included by [43], as positive values there represent a negative relationship between sexual signals and parental care.
8. Effect sizes related to aggression received that could be included by [39], as positive values there represent a positive relationship between sexual signals and aggression received.

The direction of some re-extracted effect sizes was unclear because the authors of the empirical study did not clarify the expected direction for a given sexual signal, proxy, or estimate they reported (e.g. unclear whether colour PC1 positively or negatively related to signal conspicuity, extravagance, or attractiveness in [176]). We excluded re-extracted effect sizes with an unclear proxy direction that were not matched to any reported effect sizes (unusual proxies, e.g. resting metabolic rate). For all other cases with unclear direction, whenever the re-extracted effect size was similar to their matched reported effect size(s) in magnitude (absolute difference less than 0.02) but in opposite directions, we changed the direction of that re-extracted effect size to be the same as the reported one(s) which it matched to minimise differences when information was ambiguous.

Regarding data on the relationship between condition-dependency (see Table S1 for specific proxies) and sexual signals, we expected that individuals that express sexual signals do so because they can pay the costs of developing and maintaining that trait, indicating their good condition. However, this rationale is only valid for correlational studies. When studies manipulated sexual signals of individuals, we expected that individuals whose sexual signal was enlarged or made more intense would experience increased costs and thus ultimately lower survival.