doi:10.1111/evo.13459

Irreproducible text-book "knowledge": The effects of color bands on zebra finch fitness

Daiping Wang,[1] Wolfgang Forstmeier,[1,2] Malika Ihle,[1,3] Mehdi Khadraoui,[1] Sofia Jerónimo,[1] Katrin Martin,[1] and Bart Kempenaers[1]

[1]*Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Street 7, 82319 Seewiesen, Germany*

[2]*E-mail: forstmeier@orn.mpg.de*

[3]*Current Address: Department of Entomology and Nematology, University of Florida, 1881 Natural Area Dr., Gainesville, Florida 32611*

Received September 28, 2017
Accepted February 12, 2018

Many fields of science—including behavioral ecology—currently experience a heated debate about the extent to which publication bias against null findings results in a misrepresentative scientific literature. Here, we show a case of an extreme mismatch between strong positive support for an effect in the literature and a failure to detect this effect across multiple attempts at replication. For decades, researchers working with birds have individually marked their study species with colored leg bands. For the zebra finch *Taeniopygia guttata*, a model organism in behavioral ecology, many studies over the past 35 years have reported effects of bands of certain colors on male or female attractiveness and further on behavior, physiology, life history, and fitness. Only eight of 39 publications presented exclusively null findings. Here, we analyze the results of eight experiments in which we quantified the fitness of a total of 730 color-banded individuals from four captive populations (two domesticated and two recently wild derived). This sample size exceeds the combined sample size of all 23 publications that clearly support the "color-band effect" hypothesis. We found that band color explains no variance in either male or female fitness. We also found no heterogeneity in color-band effects, arguing against both context and population specificity. Analysis of unpublished data from three other laboratories strengthens the generality of our null finding. Finally, a meta-analysis of previously published results is indicative of selective reporting and suggests that the effect size approaches zero when sample size is large. We argue that our field—and science in general—would benefit from more effective means to counter confirmation bias and publication bias.

KEY WORDS: **Color bands, fitness, null findings, publication bias, zebra finch.**

In an ideal world, scientific studies would get reported irrespective of whether findings are statistically significant (positive finding) or not (a "null result": the null hypothesis of no effect cannot be rejected). If the likelihood of reporting would be independent of the outcome of hypothesis tests, all results could be included and summarized in meta-analyses. This would allow us to obtain reliable estimates of the average size of an effect and of its variability among studies, that is, its degree of context dependence. However, current scientific practice is often far from reaching that ideal state (Begley and Ellis 2012; Collaboration 2015; Freedman et al. 2015; Baker 2016; Kousta et al. 2016; Forstmeier et al. 2017; Ihle et al. 2017). Indeed, the existing scientific literature is likely biased toward studies that report positive findings, because null results are more difficult to publish (Horton 2015; Parker et al. 2016; Forstmeier et al. 2017). Such selective reporting implies that the literature also contains a high proportion of false-positive claims (Greenwald 1975; Jennions and Moller 2002; Prinz et al. 2011; Button et al. 2013; Franco et al. 2014; Holman et al. 2016). Again, in an ideal world, claims of positive effects should motivate attempts at replication, which would then allow us to distinguish false-positive claims from true-positive effects. Unfortunately, this process of verification is hindered by

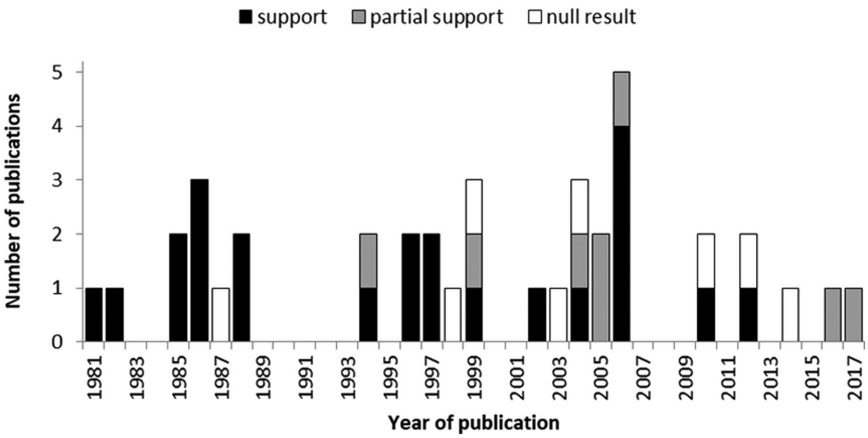

Figure 1. **Summary of publications ($n = 39$) of experiments in which male zebra finches were fitted with red versus green color bands. Shown are the number of studies and their year of publication. Studies were classified as (1) providing support ($n = 23$) for the hypothesis that red-banded males are in some way doing "better" than green-banded males, (2) providing partial support ($n = 8$) defined as showing at least some significant effect of color bands, or (3) no support ($n = 8$) defined as showing no significant effects of color bands. Year of publication is a significant predictor of whether a study was supportive or not (logistic regression, $n = 39$, $P = 0.011$).**

journals and funding agencies that prioritize novelty over solid replication (Song and Gilbody 1998; Collaboration 2015; Benjamin et al. 2017; Forstmeier et al. 2017; Szucs and Ioannidis 2017). To add insult to injury, a replication study that fails to find evidence for the originally reported effect might be difficult to publish.

Our aim is to provide an example of the general problem that the scientific literature may misrepresent reality. In behavioral ecology, the hypothesis that colorful leg bands can alter the attractiveness of male or female zebra finches (*Taeniopygia guttata*), with resulting effects on behavior, physiology, life history, and fitness, has been quite influential (Burley 1981; Burley et al. 1982; Burley 1985a; Burley 1986b; Burley 1986a; Burley 1988; Burley et al. 1994; Burley et al. 1996; Cuthill et al. 1997; Hunt et al. 1997; Gil et al. 1999; Benskin et al. 2002; Pariser et al. 2010). Zebra finches are among the most intensely studied organisms in behavioral ecology (Collins and ten Cate 1996; Riebel 2009; Griffith and Buchanan 2010; Adkins-Regan 2011), and studies of putative color-band effects not only make up a considerable part of the zebra finch literature, but also spurred and influenced the development of key concepts such as differential allocation and other maternal effects (Burley 1988), which subsequently were tested in a wide range of taxa (Sheldon 2000; Ratikainen and Kokko 2010). Color-band effects on attractiveness and other phenotypes have also been examined in various other bird species, but here the majority of studies reported null findings (Metz and Weatherhead 1991; Cristol et al. 1992; Hannon and Eason 1995; Johnsen et al. 2000; Verner et al. 2000; Cresswell et al. 2007; Roche et al. 2010 but see Brodsky 1988; Johnsen et al. 1997). Remarkably, the hypothesis of artificial color effects on attractiveness has also been studied extensively in humans. Starting with a seminal paper on the "Red-Romance Hypothesis" (Elliot and Niesta 2008), a large body of literature has accumulated showing that, for instance, wearing a red T-shirt or being shown in front of a red background strongly enhances the attractiveness of men (Elliot et al. 2010; Buechner et al. 2015) and women (Elliot and Niesta 2008; Kayser et al. 2010; Elliot and Pazda 2012; Pazda et al. 2012; Elliot et al. 2013a, 2013b; Elliot and Maier 2013; Pazda et al. 2014a; Pazda et al. 2014b). Some of these studies highlighted the parallels to the zebra finch example (Elliot et al. 2010; Elliot and Maier 2012). However, these striking results have been questioned and considered "too good to be true" in the sense that there is a clear shortage of null findings despite low statistical power (Francis 2013), and more recent studies from other laboratories report null findings despite high statistical power (Hesslinger et al. 2015; Peperkoorn et al. 2016; Lehmann and Calin-Jageman 2017).

Focusing on the zebra finch literature, we identified 39 publications reporting experimental work in which male zebra finches had been fitted with either red or green color bands, identified as having the most enhancing and most detrimental effects on male attractiveness, respectively (Burley et al. 1982). The majority (23, 59%) of these 39 publications concludes or confirms that red-banded males are in some way "superior" to green-banded males (Fig. 1; Table S1). Eight publications (21%) report that the color bands resulted in at least some significant effects (e.g., in interaction with other variables; Fig. 1; Table S1). Eight studies (21%) report that color bands had no significant effects at all (Fig. 1; Table S1). Of the latter, nearly all emphasized that low statistical power may have resulted in a false-negative conclusion (a type II statistical error), or that color-band effects may be context-specific (depending on details of the experiment) or

population-specific (depending on the origin of the birds). Only a single study (Seguin and Forstmeier 2012) questioned whether some of the previously claimed effects may in fact be absent. The temporal distribution of these 39 publications suggests that earlier studies were more often supportive, whereas more recent studies were more likely to show partial support and null results (Fig. 1).

The studies shown in Figure 1 have investigated a wide range of potential consequences of the red and green color bands, including male attractiveness to females, dominance among males, male survival and fitness, male behavior and body mass regulation, offspring sex ratio, parental effort and investment in eggs by the partner, and attractiveness as a tutor or demonstrator in social learning experiments. Most of the studies that support color-band effects report that some of the outcome variables are affected, but not others (see Schuett and Dall 2010). Nevertheless, the consensus that emerges is that red-banded males are more attractive to females than green-banded males, and in consequence achieve substantially higher reproductive success (see summary in Schuett and Dall 2010; Seguin and Forstmeier 2012). The full fitness consequences of wearing color bands have not yet been assessed in a single study, but it has been reported that red-banded males—compared to green-banded males—produced about twice as many offspring with their social partner (Burley 1986b; not accounting for extra-pair paternity), lost less paternity to extra-pair males (Burley et al. 1996), and obtained more extra-pair copulations (Burley et al. 1994). Thus, measurements of relative fitness that include parentage assignment should be most successful in capturing the sum of beneficial effects that red color bands convey and the contrasting detrimental effects of wearing green color bands.

Previous reports further suggest that bands with other colors than red or green also affect the attractiveness of zebra finches, albeit to a lesser extent (Burley et al., 1982, 1985b). However, these colors have received limited attention in experimental studies. Burley et al. (1982) reported that light blue bands were nearly as detrimental as light green (for both sexes) and that black and pink bands enhanced attractiveness and fitness components in females. Other colors appeared to be approximately neutral (Burley et al. 1982). Thus, effect sizes of different colors seem to vary more or less continuously from highly attractive, via practically neutral, to strongly detrimental (Burley 1985b).

Experimental work on zebra finches often requires marking individuals. Despite the above, most researchers appear to have avoided the use of red or green bands on males, while considering all other colors as behaviorally neutral for both sexes (Forstmeier and Birkhead 2004; Spencer and Verhulst 2007; David and Cézilly 2011).

In our previous work, we never detected any significant effects of band colors when using such potentially neutral colors (reported in Forstmeier and Birkhead 2004; Bolund et al. 2007; Forstmeier et al. 2011; Ihle et al. 2015; Wang et al. 2017a), arguing against the idea that some of these colors have at least small effects. Furthermore, earlier attempts to replicate two specific studies (included in Fig. 1) did not show any effects of red and green color bands on male behavior and body mass (Seguin and Forstmeier 2012) or on copying behavior in social learning experiments (Mora and Forstmeier 2014). Finally, our observation that zebra finch mate preferences seem predominantly individual specific rather than following a universal rule of attractiveness (Forstmeier and Birkhead 2004; Ihle et al. 2015; Wang et al. 2017a; Wang et al. 2017b) is at odds with the existence of universal band-color effects on attractiveness.

In view of the above and of the current debate about replicability of research findings (Song and Gilbody 1998; Collaboration 2015; Freedman et al. 2015; Baker 2016; Holman et al. 2016; Kousta et al. 2016; Parker et al. 2016; Benjamin et al. 2017; Forstmeier et al. 2017; Parker and Nakagawa 2017; Szucs and Ioannidis 2017), the aim of this study is to rigorously test for color-band effects on fitness in four populations of captive zebra finches (two domesticated and two recently wild-derived). For this purpose, we analyze reproductive success (fitness) as a function of band color in eight experiments, four previous experiments in which fitness of color-banded birds had been measured, but in which red and green bands had been avoided, plus four recent experiments that specifically included red and green bands. We model the fitness of males and females separately and fit band color as a random effect to reflect the working hypothesis (based on previous evidence, see above) that most if not all colors are nonneutral to some extent, and to quantify the total proportion of variance explained by this factor. To examine whether color-band effects are population- or context-specific, we also code colors differently within each of the four populations and within each of the eight experiments. An observed mismatch between our findings and the existing literature further prompted us to examine unpublished data from other laboratories and to assess publication bias in published estimates.

Materials and Methods

DATA INCLUSION CRITERIA

We included all experiments ever conducted in our laboratory in which color-banded birds raised their own offspring in communal aviaries, such that their achieved fitness (number of genetic offspring raised to independence) could be quantified. These criteria were met by eight experiments (Table 1). Three experiments were not optimally designed for the purpose of this study, but we still included them to avoid selective reporting. In experiments 3 and 4, pair bonds had already formed before the allocation of color bands (see Ihle et al. 2015). Thus, color bands could not affect

Table 1. Details of eight experiments in which fitness of zebra finches wearing bands of different colors was quantified.

Experiment	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
Population	Melbourne	Bielefeld	Bielefeld	Bielefeld	Krakow	Seewiesen	Seewiesen	Seewiesen
Origin	Wild	Wild	Wild	Wild	Domest	Domest	Domest	Domest
Housing	Outdoors	Outdoors	Outdoors	Outdoors	Outdoors	Indoors	Indoors	Indoors
Year	2016	2016	2012–2013[1]	2012	2016	2007	2009	2016
Duration (days)	93	93	2 × 86[1]	86	93	92	113	90
N males	31	29	59[1]	36	48	36	36	90
N females	29	31	59[1]	36	48	36	36	90
N aviaries	5	5	10 and 7[1]	6	8	6	6	15
Males:females per aviary	5:7 or 7:5	5:7 or 7:5	6:6 or 5:5	6:6	5:7 or 7:5	6:6	6:6	6:6
N offspring	91	58	425	133	201	144	129	259
Inbreeding *F* mean	0	0.023	0.002	0.125	0.009	0	0.121	0.110
Inbreeding *F* maximum	0	0.133	0.063	0.25	0.039	0	0.25	0.299
Colors	b, bl, lb, g, r, w, y	b, bl, lb, g, r, w, y	b, bl, lb, o, w, y	b, bl, lb, o, w, y	b, bl, lb, g, r, w, y	g-bl, g-w, r-w, r-bl, w-bl, y-bl and b, bl, o, p, w, y[2]	b, bl, o, p, w, y	bl, g, lb, o, p, r

Fitness was estimated as the number of independent offspring produced in communal aviaries, accounting for extra-pair paternity (see Methods). Birds came from four populations, two recently wild-derived (wild) and two domesticated (domest). They were housed either in semi-outdoor aviaries with natural and artificial light, or indoors under artificial light only. The year of study and the duration of the breeding period (period during which birds were allowed to lay eggs, excluding the time allowed for raising offspring) is indicated. The total number of individual males and females and their distribution among aviaries is shown, as well as the total number of offspring that were raised to 35 days of age. The mean and maximum inbreeding coefficient *F* of all adults is also shown. Abbreviations for color bands used: b = dark blue, bl = black, lb = light blue, g = green, r = red, w = white, y = yellow, o = orange, p = pink; two-colored striped bands in Exp. 6 are explained in the footnote.

[1]Fifty-nine males and 59 females bred for 86 days in 2012 in 10 aviaries; a subset of 41 males and 41 females bred a second time for 86 days in 2013 in seven aviaries with different color bands (by swapping colors, see Methods for details).

[2]The birds were banded twice: during the first 14 days of the experiment, birds received striped color bands (green-black, green-white, red-white, red-black, white-black, and yellow-black) and from day 15 onwards they received the usual uniform color bands.

pair formation, but they could still affect fitness via differential allocation (Burley 1988) and via effects on extra-pair paternity gain (Burley et al. 1994) and paternity loss in the own brood (Burley et al. 1996). In these experiments, the effect of color bands on fitness may thus be smaller than in other experiments. In experiment 6, individuals were color-banded with one set of bands from day One to 14, primarily affecting pair formation, and then received a different set of color bands, which might have affected differential allocation and paternity (in total, the egg-laying period lasted 92 days plus about 50 days for chick rearing). To deal with this, we carried out two analyses: one using the initial color and one using the final color as a predictor. We also analyze band-color effects on fitness in a reduced dataset (excluding experiments 3, 4, and 6).

GENERAL PROCEDURES

Details of the eight experiments are summarized in Table 1. They comprise work on four different captive populations, two of which are domesticated and two of which are recently wild-derived (for details see Supplementary Information). Breeding took place in two types of aviaries: indoor aviaries with artificial light (see Wang et al. 2017a) or semioutdoor aviaries that include natural light (Ihle et al. 2015; Jerónimo et al. 2018). The aviaries initially contained 12 adult birds, usually six females and six males (but in 14 out of 68 experimental aviaries one individual died during the experiment, in six aviaries two individuals died, and in two aviaries three individuals died). However, in three experiments a sex-ratio bias was created with either seven females to five males, or five females to seven males. Hence, we always include the initial adult sex-ratio (i.e., proportion of males: 0.417, 0.5, or 0.583) as a fixed effect in our analyses of reproductive success. In three experiments individuals varied substantially in their level of inbreeding, so in all analyses, we also control for an individual's inbreeding coefficient (calculated using Pedigree Viewer 6.4a, Kinghorn and Kinghorn 2010). Finally, the experiments lasted

between 86 and 113 days, whereby all eggs laid within this period were allowed to be reared to independence, usually requiring another seven weeks. Thus, we include experimental duration as a fixed effect in analyses of reproductive success.

Reproductive success was quantified as the number of genetic offspring that reached 35 days of age (usually regarded as the age of independence, Sossinka 1980; Ihle et al. 2015). Genetic parentage assignment was based on data from 12 to 16 microsatellite markers (see Wang et al. 2017b), which allows for a practically error-free assignment as confirmed by SNP genotyping (Backström et al. 2010; Knief et al. 2017). Reproductive success was calculated for all birds that were present at the start of the experiment ($N_{\text{total}} = 367$ males and 367 females), including the ones that later died ($N_{\text{died}} = 10$ males and 22 females), with one exception. In experiment 3, designed to measure the fitness of prearranged pairs (see Ihle et al. 2015), two birds were removed when their partner died and these were excluded from the analysis. In the same experiment, a subset of 41 males and 41 females (out of 59 males and 59 females) were measured for fitness twice (see Table 1), while wearing different color bands. We included these repeated measures of reproductive success in the analyses accounting for individual identity as a random effect. Hence, in total we analyzed reproductive success based on 1440 offspring raised to independence by 365 individual males and 365 individual females from a total of 406 male breeding seasons and 406 female breeding seasons.

COLOR BANDS

Color bands (size XCS for domesticated populations and XF for recently wild-derived populations, obtained from A. C. Hughes, Hampton Hill, U.K., maximum nine different colors) were used for individual identification, such that each color was used only once per sex and aviary. Each bird received two bands of the same color, one on each leg. For optimal visibility, the color band was placed below the metal band (anodized orange) on the right leg. Colors were assigned to individuals using the random-number function in Excel. Birds could choose their partner among the available individuals, except in experiments 3 and 4, where pairs had been formed prior to the start of breeding (see above). In those experiments, colors were randomly assigned to pairs rather than to individuals such that the members of a pair wore the same color (unless they divorced and repaired). In experiment 6, where colors were changed after 14 days, the assignment of initial bands was random, but the new set of bands were again allocated to pairs, whereby members of a pair were given different but randomly predefined colors (see Supporting Information for more detail). The color bands used during the first 14 days of experiment 6 differed markedly from the ones we used otherwise: they were two-colored ("striped") rather than uniform, with one color in the top half and the other in the bottom half (see Table 1). Thus, in the analysis, the variable "color band" has up to 15 categories: six striped color combinations plus nine uniform colors.

STATISTICAL ANALYSES

For illustrative purposes only, we calculated relative fitness of individuals within each aviary scaled to an average of unity, and we show the average relative fitness of birds of a given band color for each experiment (separately for each sex).

For statistical analyses, we used linear mixed-effect models (lme4 package, Bates et al. 2015; in R 3.2.3, R Core Team 2015) to investigate the effect of color bands on individual reproductive success in each sex across all experiments and populations. The number of independent offspring produced per breeding season by each individual was square-root transformed to reduce the deviation from normality and was modelled as a Gaussian trait in separate models for males and females. Individual identity (365 levels), aviary identity (68 levels), experiment identity (8 levels), and population identity (4 levels) were always included as random effects. Band color was also included as a random effect, reflecting the working hypothesis that all colors can have some effect on attractiveness, with red and green presumably having the strongest effect in males. As described above, in version 1, we fitted the initial band colors including the striped bands (used in experiment 6) as a random effect (15 levels of color), whereas in version 2, we fitted the final band colors (nine levels of uniform color). To test the idea that color-band effects may be specific to the population or specific to the experiment, we also coded colors uniquely within populations (31 levels) and within experiments (51 levels) and fitted these as random effects. As fixed effects we controlled for the adult sex ratio within the aviary, the duration of the breeding season in days, and the individual's inbreeding coefficient, as explained above. To examine the hypothesis that red and green bands exhibit specific effects on male fitness, we also fitted "red versus green band" as a fixed covariate. We coded red as +0.5, green as –0.5, and all other colors as 0, so that the regression slope quantifies the increase in number of offspring sired (square-root transformed) from green to red.

RELATING OUR RESULTS TO EXPECTATIONS FROM THE LITERATURE

To illustrate how our results relate to expectations from the literature (see Introduction), we plot the mean relative fitness of individuals with a given band color over an arbitrary "attractiveness rank" derived from the literature (Burley et al. 1982). To do this, we classified colors as either attractive (scored as +0.5: red for males, black and pink for females), neutral (scored as 0: orange and red for females, pink, orange, and black for males), or unattractive (scored as -0.5: light blue and green for both sexes). This quantification allowed us to add "attractiveness rank" as another covariate to the mixed models described in the previous

section. In an alternative version of analysis, we post hoc lumped the striped color bands containing green or red with the uniform green or red bands (red–black and red–white coded as red; green–black and green–white coded as green), that is, we categorized them using the colors with the strongest expected effects.

ANALYSIS OF UNPUBLISHED DATA FROM OTHER LABORATORIES

In 2001, Nikolaus von Engelhardt initiated a replication study of the presumed effect of red and green color bands on offspring sex ratio (Burley, 1981, 1986a). This project was carried out collaboratively across three laboratories (at the Universities of Groningen, Bielefeld, and Melbourne), but the results were only published in a Ph.D. thesis (von Engelhardt 2004). Under the kind permission of von Engelhardt and his collaborators, we used their summarized data on offspring production (Table 2.1 on page 21 of von Engelhardt 2004) to calculate the relative fitness of males wearing different color bands (red, orange, or green, from the same source: A. C. Hughes, Hampton Hill, U.K.). Their experiments closely followed the design described in Burley (1986a,b): aviaries contained 24 males and 24 females, males received two bands of the same color (eight males per color), all females received two orange bands. Four such aviaries were set up in Groningen (domesticated population), one in Bielefeld (recently wild-derived population), and one in Melbourne (recently wild-derived population). Over a period of three months, the 144 males produced a total of 157 offspring (surviving young to sexual maturity) in their own nest. Thus, the measure of reproductive success is based on social parentage (as in Burley 1986b) rather than on genetic parentage assignment.

To analyze the summarized data (number of offspring produced, averaged among eight males of the same color, with three colors times six aviaries resulting in 18 mean values), we ran a mixed effect model with the mean number of offspring (square-root transformed) as the dependent variable, and aviary ($n = 6$) and population ($n = 3$) as random effects to account for nonindependence. As the only fixed effect we fitted "attractiveness rank" as defined in the previous paragraph (red = 0.5, orange = 0, green = –0.5). Although this model is based on few datapoints, the slope estimate for "attractiveness rank" can be compared to the estimate from our own populations.

EXTRACTION OF EFFECT SIZE ESTIMATES FROM THE LITERATURE

From the 39 publications shown in Figure 1, we extracted estimates of effect size of males wearing green versus red color bands (main effects only, without interactions). We classified the diverse dependent variables into two groups: those related to male–male competition (male body mass, male dominance) and those putatively mediated by female choice (e.g., approach times in a choice test, copulation rates, measures of parental effort, yolk hormone concentrations, offspring sex ratio). Band-color effects on metric traits were quantified as Cohen's D (Cohen 1988) with measures of SD sometimes approximated from reported ranges or from related publications (see Supporting Information File). Effects on binomial traits such as sex ratio were usually expressed as odds ratios and then converted to Cohen's D using a website resource from Lenhard and Lenhard (2016). In total, we obtained 141 effect size estimates with their respective sample size N (see Supporting Information File). We acknowledge that this data extraction contains elements of arbitrariness (e.g., exclusion of practically redundant estimates, or quantification of offspring sex ratio at the level of the individual male or at the individual offspring level) but all information is given in the Supporting Information.

FUNNEL PLOTTING AND ANALYSIS OF AVERAGE EFFECT SIZE AND STATISTICAL POWER

We first plotted all 141 estimates of effect size (Cohen's D) over their respective sample size (inverse of the square-root of sample size, $N^{-0.5}$) and tested for asymmetry in this funnel plot using the R Package "meta" (Schwarzer and Schwarzer 2017). We also tested for asymmetry separately for estimates related to female choice ($N = 129$). Estimates related to male–male competition ($N = 12$) had been summarized previously in Seguin and Forstmeier (2012) and were too few for meaningful analysis. In light of a dispute about the best methods (see Tang and Liu 2000; Sterne and Egger 2001), we also used the R Package "metafor" (Viechtbauer 2010; Nakagawa et al. 2015) to test for asymmetry in a funnel plot of effect size over its SE (rather than over $N^{-0.5}$). The two methods differ in their definition of precision (the former depends on N only, the latter depends on N and effect size), and we apply both methods to examine the robustness of our conclusion. The "metafor" package was also used to quantify heterogeneity in the 141 observed effect sizes.

To analyze variation in effect sizes, we specified a mixed effect model with Cohen's D as a Gaussian dependent trait, weighted by sample size (i.e., by the square root of $N - 3$). Trait category (competition or choice) was entered as a fixed effect, year of publication as a continuous covariate, and population identity (16 levels) and identity of the research group (13 levels) as random effects. The two random effects were strongly aliased, with only three research groups having data from two or three study populations. This means that it is not meaningful to try separating the two random effects, but both were kept in the model to control for the nonindependence of datapoints. Random effect estimates were examined for outliers, and outliers were subjected to separate tests for average effect size and for asymmetry in the funnel plot. Making the assumption that all reported effect sizes correspond to true effects, we calculated the statistical power of published tests

Table 2. **Linear mixed model explaining variation in reproductive success (square-root transformed number of independent offspring per breeding season) of 365 female zebra finches ($N = 406$ female breeding seasons).**

	Estimate ($\beta \pm$ SE)	t	P
Random effects:			
Female ID ($n = 365$)	0.468		
Aviary ($n = 68$)	0.000		
Band color ($n = 15$ or 9)[1]	0.000		
Experiment ($n = 8$)	0.042		
Population ($n = 4$)	0.000		
Residual	0.557		
Fixed effects:			
Intercept	1.538 ± 0.092	16.7	-
Adult sex ratio	1.203 ± 1.189	1.01	0.31
Duration of breeding season (d)	0.006 ± 0.012	0.48	0.63
Inbreeding coefficient	-3.644 ± 0.748	-4.87	<0.0001

For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

[1]Two versions of the model using different data from Experiment 6. Version 1 included individuals with the original bands (15 band colors, including striped bands); version 2 included individuals with replaced uniformly colored bands (nine band colors). Note that in both model versions the variance component associated with "band color" equaled zero, so the other estimates are not affected by model version.

for finding the reported effect size using the software G*Power 3.0.10 (Faul et al. 2009).

Results

FACTORS EXPLAINING VARIATION IN REPRODUCTIVE SUCCESS

Variation in reproductive success was largely explained by the same factors in females (Table 2) and males (Table 3). Reproductive success was individually repeatable in both sexes (female identity explained 44% of the variance, male identity explained 33% of the variance). However, these estimates should be considered with caution, because birds were measured repeatedly only in experiment 3. Reproductive success varied slightly between the eight experiments (explaining 4% of variance in females, 3% in males), but did not vary systematically between the four populations or between the 68 experimental aviaries (variance components equaled zero). Reproductive success declined strongly with the individual's inbreeding coefficient, with a similar slope in females and males (Tables 2 and 3). As expected, the effect of the adult sex ratio in the aviary differed between the sexes. With an increasing proportion of males, female reproductive success nonsignificantly increased (Table 2), while male reproductive success significantly decreased (Table 3). Finally, the duration of the breeding season (see Table 1) had little effect on female and male reproductive success (estimates are both positive, but small and nonsignificant, Tables 2 and 3).

GENERAL COLOR-BAND EFFECTS ON REPRODUCTIVE SUCCESS

Reproductive success appeared to vary randomly with regard to band color in both females (Fig. 2) and males (Fig. 3). Indeed, band color as a random effect explained 0% variance in female (Table 2) and in male (Table 3) reproductive success, irrespective of how we classified colors in experiment 6 (see Tables 2 and 3 and Methods for details). Analyses of the reduced dataset (excluding the suboptimally designed experiments 3, 4, and 6) led to identical conclusions (see Supporting Information Tables S4 and S5).

POPULATION- OR CONTEXT-SPECIFIC BAND-COLOR EFFECTS

To examine whether band colors had population-specific effects on reproductive success, we recoded colors within populations (31 color-population combinations used, Table 1; yielding on average 13.1 measures of reproductive success per level for each sex). This random effect explained 0.17% of the variance in female reproductive success ($P = 0.49$) and 0% of the variance in male reproductive success ($P > 0.5$).

Similarly, to estimate context-specific band-color effects, we recoded colors within experiments (51 color-experiment combinations used, Table 1; on average eight measures of reproductive success per level for each sex). The variance component for this random effect was zero for both females and males. Changing to the other version of analysis for experiment 6 led to the same conclusions (the variance components were also zero or close to zero).

Table 3. Linear mixed model explaining the variation in reproductive success (square-root transformed number of independent offspring sired per breeding season) of 365 male zebra finches ($N = 406$ male breeding seasons).

	Estimate ($\beta \pm$ SE)	t	P
Random effects:			
Male ID ($n = 365$)	0.411		
Aviary ($n = 68$)	0.000		
Band color ($n = 15$ or 9)[1]	0.000		
Experiment ($n = 8$)	0.036		
Population ($n = 4$)	0.000		
Residual	0.786		
Fixed effects:			
Intercept	1.478 ± 0.090	16.4	-
Adult sex ratio	-3.347 ± 1.282	-2.61	0.009
Duration of breeding season	0.005 ± 0.012	0.42	0.67
Inbreeding coefficient	-3.696 ± 0.800	-4.62	<0.0001
Red versus green band[2]	-0.017 ± 0.231	-0.08	0.94

For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

[1]Two versions of the model using different data from experiment 6. Version 1 included individuals with the original bands (15 band colors, including striped bands); version 2 included individuals with replaced uniformly colored bands (nine band colors). Note that in both model versions the variance component associated with "band color" equaled zero, so the other estimates are not affected by model version.

[2]The reported effect is for version 1 of the model (red-striped pooled with red, green-striped pooled with green). In model version 2, the estimate changes to -0.299 ± 0.269, $t = -1.11$, $P = 0.27$.

CONSISTENCY WITH PREVIOUS FINDINGS

Figure 4 illustrates the relationship between average relative fitness for each band color and their proposed attractiveness rank based on the literature (see Methods). In version 1 of our analysis, we lumped the striped color bands used in the first two weeks of experiment 6 into the categories of red and green (see Methods). This was done post hoc to allow maximum support for the hypothesis, given the observation that males with red-striped bands achieved higher fitness than males with green-striped bands (see experiment 6(1) in Fig. 3; two-sample t-test, $N_{red} = 12$ males, $N_{green} = 12$ males, $t_{22} = 1.77$, $P = 0.091$). Overall, in this version of analysis, red-banded males had a slightly higher average relative fitness than green-banded males (Fig. 4, bottom left). However, in a mixed-effect model that also accounts for the effects of inbreeding and other covariates, the estimated number of offspring produced by red-banded and green-banded males did not differ (negative slope of -0.017 ± 0.231, $P = 0.94$, Table 3). Under version 2 of the analysis (using the data from experiment "6(2)" with only uniformly colored bands), if anything, red-banded males tended to perform worse (negative slope of -0.299 ± 0.269, $p = 0.27$, Table 3). Corresponding models using the attractiveness rank as shown in Figure 4 yielded weakly negative slopes that are opposite to expectations (version 1: -0.060 ± 0.226, $P = 0.79$, Table S8; version 2: -0.266 ± 0.220, $P = 0.23$, Table S9). For females the corresponding slopes were weakly positive, yet far from significant (version 1: 0.043 ± 0.162, $P = 0.79$, Table S6; version 2: 0.098 ± 0.198, $P = 0.62$, Table S7).

UNPUBLISHED DATA FROM OTHER LABORATORIES

Based on data from von Engelhardt (2004), the observed relative fitness of males with red, orange, and green color bands was not consistent with expectations from the literature in any of the three captive populations (Fig. 5). Similarly, a mixed-effect model with aviary ($n = 6$) and population ($n = 3$) as random effects showed that "attractiveness rank" was, if anything, negatively related to social reproductive success (slope: -0.555 ± 0.568, $P = 0.33$).

ANALYSIS OF PUBLISHED EFFECTS

The effect size estimates extracted from the published literature ($N = 141$) were significantly related to sample size (test for asymmetry in the funnel plot: $P = 0.019$; based on "meta" Schwarzer and Schwarzer 2017). The 129 estimates related to effects of female choice showed a strong asymmetry ($P = 0.009$; gray line Fig. 6), whereby effect size reached zero at highest sample sizes. When effect sizes were plotted over their respective SEs, the asymmetry of the funnel plot was even more pronounced ($P = 0.0017$; based on "metafor" Viechtbauer 2010; Nakagawa et al. 2015). Heterogeneity in the observed effect sizes was high (total heterogeneity/total variability = 73%, $P < 0.0001$).

Variation in effect sizes was not explained by population ID (random effect with $N = 16$ levels, variance = 0), but partly by

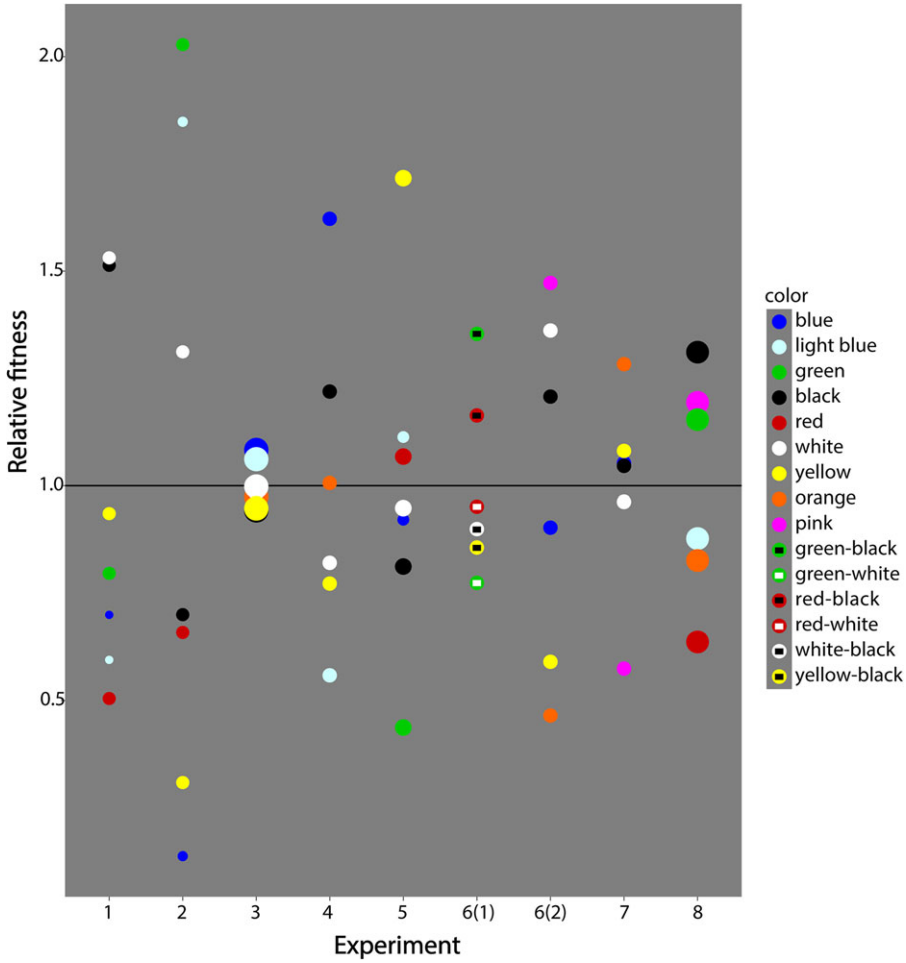

Figure 2. **Mean relative fitness of female zebra finches by band color for each of eight experiments. Each dot represents the average relative fitness (number of independent offspring) of all females with that color band. The size of the dots reflects sample size (number of females ranging from 2 to 17, most frequently 6; for details see Table S2). Relative fitness is calculated to have a mean of one in each experiment (horizontal black line). Experiment number is indicated (see Table 1 and Methods for details). In experiment 6, females wore bicolored striped bands during the first two weeks of the experiment (6(1)), which were then exchanged for the regular uniformly colored bands (6(2)). Relative fitness was analyzed for the initial color bands (version 1) and for the final color bands (version 2).**

research group ID (random effect, $N = 13$ levels, 4.4% of variance). Note, however, that these two effects cannot be distinguished with any confidence because the levels are strongly aliased. The effect of research group was mostly driven by a single group (the one where the effect had initially been discovered), who reported fivefold larger effects ($d = 1.09 \pm 0.22$, $t = 4.9$, $P = 10^{-6}$) than all other groups combined ($d = 0.22 \pm 0.08$, $t = 2.7$, $P = 0.008$; Fig. 6). Furthermore, the asymmetry in the funnel plot became nonsignificant when data from this research group ($N = 22$) were taken out ($P = 0.12$, $N = 107$; Fig. 6). Finally, we note that all 22 published estimates from this research group were statistically significant ($P < 0.05$) with an average power for the observed large effects equaling 0.79. This implies that a nonsignificant result is expected in four to five out of the 22 tests and that the combined probability of all 22 tests turning out significant is $P = 0.002$ (product of all power estimates).

Discussion

A comprehensive analysis of all available data on fitness consequences of color bands from our laboratory combined with unpublished data from another initiative to replicate studies reporting color-band effects has yielded a clear conclusion: we found no support for the previously claimed effect. Color of the bands was not associated with male or female fitness across a total of 11 experiments, seven captive populations, and four laboratories (see Figs. 4 and 5). A variance component analysis revealed that band color explained none of the observed variance in reproductive success, irrespective of whether one assumes these effects to be universal (Tables 2 and 3) or whether the effects were allowed to vary between populations or between experiments (i.e., context specificity, see Results). This means that we and other laboratories cannot robustly reproduce effects for which the literature appears

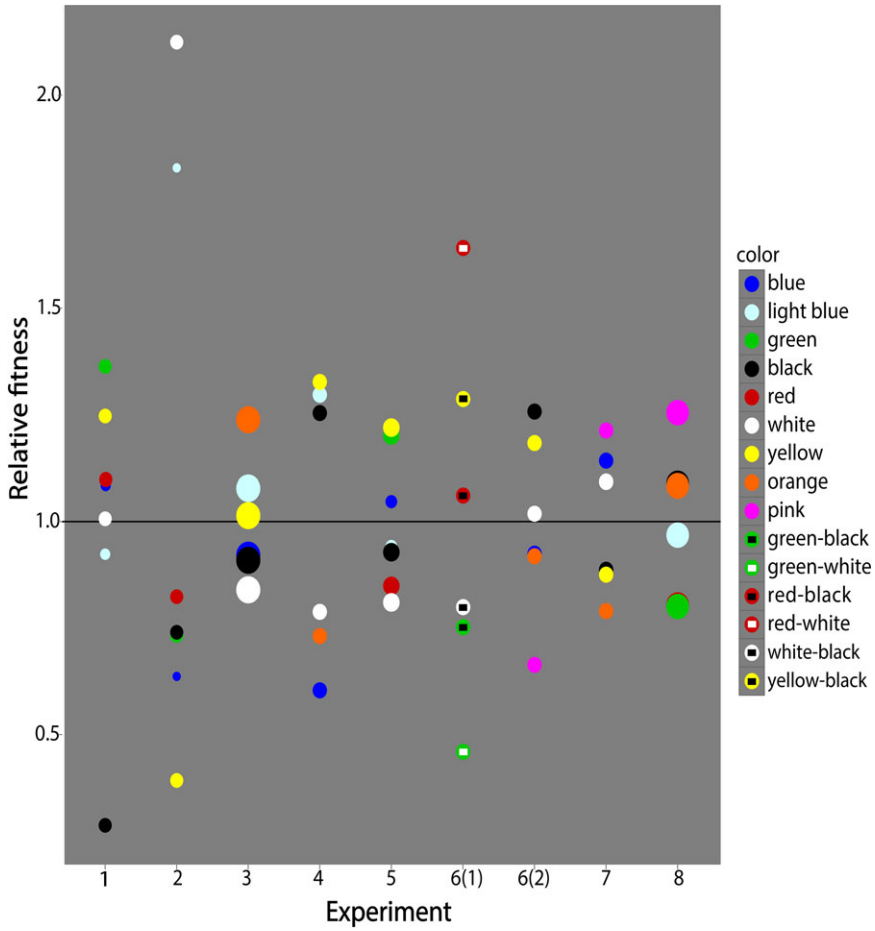

Figure 3. **Mean relative fitness of male zebra finches by band color for each of eight experiments. Each dot represents the average relative fitness (number of independent offspring sired) of all males with that color band. The size of the dots reflects sample size (number of males ranging from 2 to 17, most frequently 6; for details see Table S3). Relative fitness is calculated to have a mean of one in each experiment (horizontal black line). Experiment number is indicated (see Table 1 and Methods for details). In experiment 6, males wore bicolored striped bands during the first two weeks of the experiment (6(1)), which were then exchanged for the regular uniformly colored bands (6(2)). Relative fitness was analyzed for the initial color bands (version 1) and for the final color bands (version 2).**

to show strong evidence (see Fig. 1). This comprises both an attempt at exact replication of a specific experiment across different laboratories (data from von Engelhardt 2004) and attempts of conceptual replication (summation of all fitness-relevant effects, including within- and extra-pair success, in our experiments).

The results reported here contradict the hypothesis that all band colors have at least some effect on fitness. They also contradict the hypothesis of context- or population-specificity of effects, which often gets invoked as a post hoc explanation after a failure to confirm previous findings (e.g., Jennions 1998; Schuett and Dall 2010). This can be interpreted as an example where the existing scientific literature is biased and fails to adequately describe the biological reality. Interestingly, the data compiled by von Engelhardt (Fig. 5) remain unpublished (except in a PhD thesis) and several other research groups have carried out experiments using red and green color bands on zebra finches with null findings that remain unpublished (Jonathan Wright, Tim Birkhead, pers. comm.). Some studies that produced only null results have been published, albeit in lower impact journals (e.g., Nakagawa and Waas 2004; Schuett and Dall 2010). These studies may be perceived as reporting type II errors arising from limited power. However, in the light of our findings, the studies showing (partial) support may have reported type I errors instead. This is particularly likely in the studies showing partial support, because of multiple testing of hypotheses that were derived from the data rather than specified a priori (e.g., interaction terms). Finally, the conclusion from the literature that the effects of color bands are pervasive and hence of great biological relevance, ranging from effects on attractiveness and behavior to physiology and life history, can also be questioned. Few studies have demonstrated simultaneous effects on multiple traits, and single positive findings could also arise from multiple testing of various dependent variables

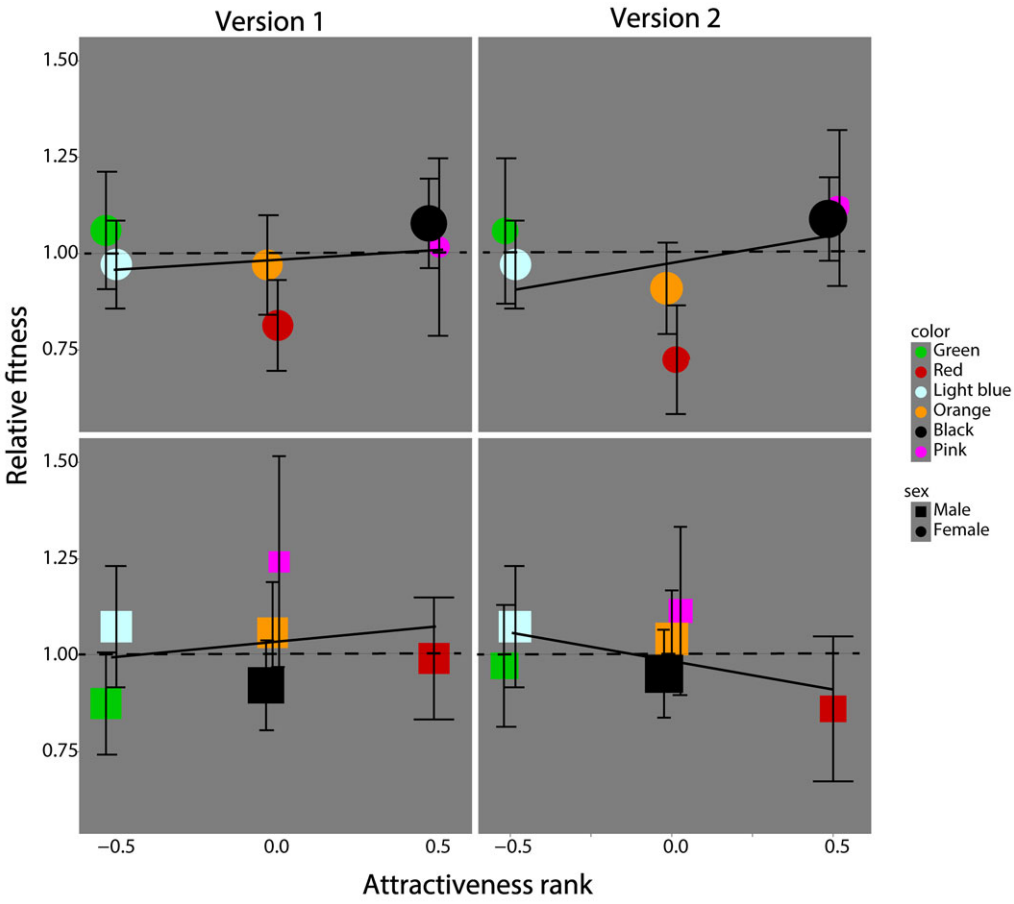

Figure 4. Regression of mean relative fitness of female (top row) and male zebra finches (bottom row) across all eight experiments as a function of the suggested attractiveness rank of each band color (based on the literature, see Introduction and Methods). Attractive colors were coded as +0.5, unattractive colors as −0.5, and neutral colors as zero. Each dot represents the average relative fitness (number of independent offspring, based on parentage analysis) of all females or all males with that color band (*N* ranging from 21 to 68, indicated by dot size). Error bars (SE) were calculated across individuals (irrespective of experiment). Scatter was introduced to the *x*-axis to increase visibility of SEs. The horizontal black dashed line indicates the mean fitness of one. In version 1 of the analysis (left panel), striped color bands containing green or red from experiment 6(1) were lumped with the categories "green" and "red". Version 2 of the analysis instead includes the uniformly colored bands from experiment 6(2). Ordinary least square regression lines (black continuous lines) are indicated for illustrative purposes only (not accounting for other effects or variation in sample size). Note that a positive slope with a twofold higher relative fitness of attractive compared to unattractive colors was expected based on effect sizes from the literature (Burley et al., 1982, 1994, 1996; Burley 1986b;).

and from selective reporting of significant effects. Future studies may want to use preregistration of hypotheses and methods (Forstmeier 2017) to ensure complete reporting of all variables that were of genuine interest (before the start of data mining) and to guard against post hoc modification of analysis strategy that can inflate effect size estimates (Simmons et al. 2011; Forstmeier et al. 2017).

Our analysis of published effect size estimates in relation to sample size strongly suggests publication bias (selective reporting), because the mean effect size approaches zero when sample size is large (Fig. 6). Note, however, that part of this apparent decline in effect size with sample size could result from heterogeneity in measurement error across estimates. For instance, one study may have reported treatment effects on offspring sex ratio at the level of the individual offspring (large number of offspring, but high noise component in the individual binomial outcome), whereas another study may have reported effects on the average proportion of sons for the color-banded fathers (smaller number of fathers, but sex ratio measured more accurately). Because effect sizes are quantified relative to the between-individual SD, they may be larger when individual values are measured with greater precision (i.e., at lower sample sizes in the above example). Nevertheless, when the true effect size >0 (true biological effect), we do not expect effect sizes to converge to zero at larger sample sizes, as suggested by the regression lines in Figure 6.

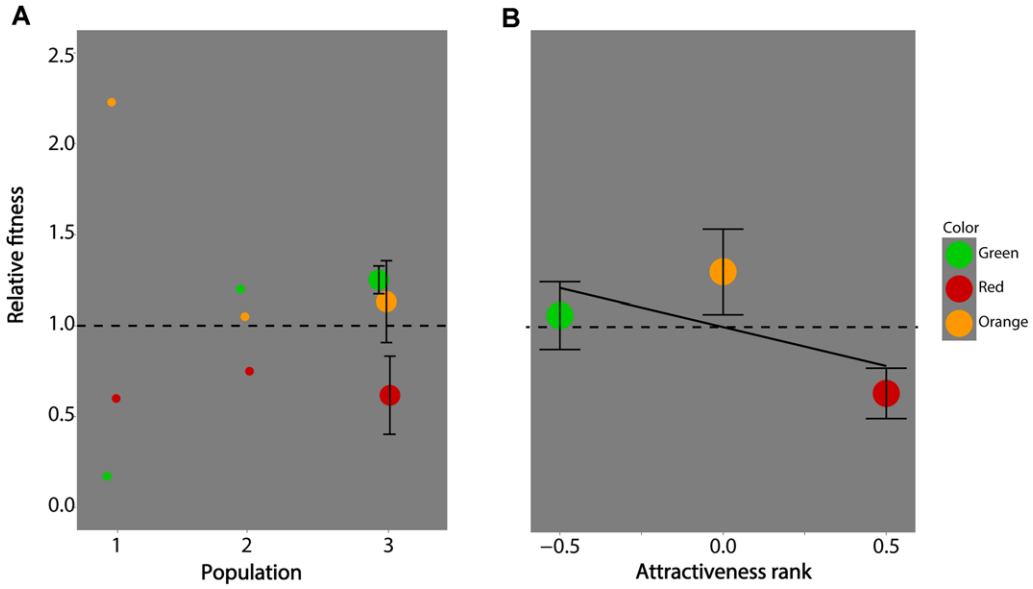

Figure 5. Summary of results from other laboratories (von Engelhardt 2004). (A) Mean relative fitness of male zebra finches as a function of band color in three captive populations (1: data collected by K. Witte in Bielefeld, (2) R. Zann in Melbourne, (3) N. von Engelhardt in Groningen). Each dot represents the average relative fitness (number of independent offspring from the own nest, not based on parentage analysis) of all males with that color band. The size of dots reflects sample size (8 or 36 males coming from one or four experimental aviaries, respectively). Because data are available only at the level of experimental aviaries, SEs are only indicated for estimates from population 3 and should be interpreted cautiously (since $n = 4$). Scatter in the *x*-axis was introduced to increase visibility of SEs. (B) Regression of mean relative fitness of male zebra finches across three populations as a function of the suggested attractiveness rank of each band color (based on the literature, see Introduction and Methods). In both panels, the mean fitness of one is indicated by a horizontal dashed black line. In (B) the continuous black represents the ordinary least square regression line (for illustrative purposes only, not accounting for other effects or variation in sample size). Here, SEs are calculated from $n = 6$ aviaries.

Underreporting of nonsignificant effects appears most pronounced (exceeding chance levels) for the research group that first described the color-band effects. For most research groups, it is plausible that statistically significant chance findings (type 1 errors) were more likely to get reported than nonsignificant test outcomes. This source of bias may explain the overall significant, yet small, main effect from published analyses from other research groups (light blue line in Fig. 6b), which we cannot reproduce in our study (Figs. 4 and 5).

Null findings are typically hard to publish because they are perceived as less informative than significant results (the so-called "Aversion to the Null", Ferguson and Heene 2012). Null results are often discarded because (1) they might represent type II errors due to limited statistical power, (2) they might arise from a failure to apply the treatments correctly, and (3) they might indicate some context-specificity of effects that is difficult to capture. In the case of zebra finch color-band effects on fitness, none of the three arguments appears convincing. (1) Statistical power: the 23 supportive publications (as categorized in Fig. 1) have been based on a total of 728 treated individuals (mean of 35 individuals per study in 21 different experiments; Table S1). For comparison, our analyses are based on 812 informative datapoints from 730 different individuals (Tables 2 and 3). Hence, for any effect size that reaches statistical significance based on 35 individuals, we have an effective statistical power of one. (2) Issues with the experimental treatment: the experimental treatment could have failed if birds were unable to perceive the band colors (e.g., due to different conditions between artificial and natural light that might affect the perception of UV), or if the birds did not show their natural behavior (e.g., due to stress). Positive findings on color-band effects have been reported from environments with artificial and natural light, and both settings were about equally represented in our experiments (Table 1). Further, none of the color bands reflects in the UV range (McGraw et al. 1999). Given that the birds bred and successfully raised offspring in all experiments, it is hard to argue that they were stressed or not showing natural behavior. (3) Context-specificity: our analyses show no heterogeneity in outcomes with regard to band color (see Tables 2 and 3 and Results). This means that the scatter of datapoints in Figures 2 and 3 correspond to the amount of noise expected under randomness. This observation argues against the idea that at least some colors exhibited effects under some conditions (or in some populations). Furthermore, our analysis of effect sizes from published data found no evidence for population-specificity of effects. Context-specificity is often

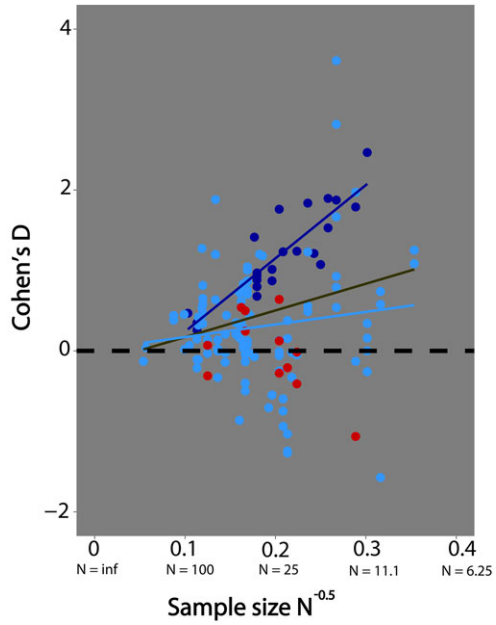

Figure 6. Funnel plot showing published effect size estimates (Cohen's *D* for red vs. green color bands, $n = 141$) in relation to their sample size. The *x*-axis shows sample size $N^{-0.5}$, where *N* is the total number of males (red plus green), or offspring (of red plus green males), or females (in choice tests). Red dots show effects related to male–male competition ($n = 12$), blue dots (light or dark) show effects related to female choice ($n = 129$); dark-blue dots represent estimates from the research group that first described the color-band effects (Burley 1981; $n = 22$). The regression lines show how effect size changes with sample size for all effects related to female choice (gray line: n = 129, $P = 0.009$), for effects from the initial group (dark-blue line: n = 22, $P = 10^{-5}$) and for effects from all other research groups (light-blue line: n = 107, $P = 0.12$). The dashed black line marks the zero.

invoked when the results of studies diverge, or concluded based on statistically significant heterogeneity in effect sizes observed in meta-analyses summarizing published data. However, such heterogeneity can also arise from biases in analysis and reporting, thereby making it hard—if not impossible—to separate biological heterogeneity from researcher-driven heterogeneity (Ferguson and Heene 2012; Forstmeier et al. 2017).

Our experiments and those initiated by von Engelhardt cannot rule out that true color-band effects have occurred at some time in some place. However, they do show that such effects are typically absent. Isolated cases of apparent, but weak support (see results of experiment 6(1) in Fig. 3, and analysis in Results) should be regarded with skepticism, because of both confirmation and attention bias (more attention given toward significant results, Forstmeier et al. 2017). We conclude that the current evidence does not support the hypothesis that color bands have pervasive effects on attractiveness, behavior, physiology, and life history of zebra finches. The current evidence rather suggests that wearing color bands is of no biological relevance to zebra finches.

The absence of universal band-color preferences corroborates the conclusions from recent work suggesting that species with socially monogamous mating systems have evolved individualistic rather than uniform mating preferences. In monogamous systems, strong preferences for attractive individuals may not be favored by selection, because the costs of competition can outweigh the benefits of choosiness (Dechaume-Moncharmont et al. 2016; Wang et al. 2017a). Instead, individualistic preferences for traits that affect behavioral compatibility and lead to optimal biparental brood care may prevail (Ihle et al. 2015). Whether zebra finches have evolved individualistic preferences that lead to repeatable between-individual differences in band color preferences (see Forstmeier and Birkhead 2004; Song et al. 2017) might be an interesting avenue for future research.

AUTHORS CONTRIBUTION

WF, DW, and BK conceived the project. DW and WF analyzed the data. BK, DW, and WF wrote the manuscript with contribution of MI and MK. Data collection was carried out by MK, SJ (experiments 1, 2, and 5), MI (experiments 3 and 4), KM (experiment 6), and DW (experiment 8).

ACKNOWLEDGMENTS

We thank N. von Engelhardt, K. Witte, the late R. Zann, T. G. G. Groothuis, F. Weissing, the late S. Daan, C. Dijkstra, and T. Fawcett for providing their data for use in this study; S. Janker, J. Schreiber, and T. Aronson for help with breeding experiments; M. Schneider for molecular work, L. J. Eberhart-Phillips for data visualization and S. Bauer, A. Kortner, J. Didsbury, and P. Neubauer for animal care. We also thank S. Nakagawa for help with the meta-analysis, M. Noor, J. Hadfield, and two anonymous reviewers for their constructive comments on the manuscript. This work was supported by the Max Planck Society (to BK) and the China Scholarship Council (CSC; stipend to DW).

DATA ARCHIVING

The doi of our data is: https://doi.org/10.5061/dryad.1hb154s.

LITERATURE CITED

Adkins-Regan, E. 2011. Neuroendocrine contributions to sexual partner preference in birds. Front Neuroendocrinol. 32:155–163.

Backström, N., W. Forstmeier, H. Schielzeth, H. Mellenius, K. Nam, E. Bolund, M. T. Webster, T. Öst, M. Schneider, and B. Kempenaers. 2010. The recombination landscape of the zebra finch Taeniopygia guttata genome. Genome Res. 20:485–495.

Baker, M. 2016. Is there a reproducibility crisis? Nature 533:452–454.

Bates, D., M. Machler, B. M. Bolker, and S. C. Walker. 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67:1–48.

Begley, C. G., and L. M. Ellis. 2012. Raise standards for preclinical cancer research. Nature 483:531–533.

Benjamin, D., D. R. Mandel, and J. Kimmelman. 2017. Can cancer researchers accurately judge whether preclinical reports will reproduce? PLoS Biol. 15:e2002212.

Benskin, C. M. W. H., N. I. Mann, R. F. Lachlan, and P. J. B. Slater. 2002. Social learning directs feeding preferences in the zebra finch, Taeniopygia guttata. Anim. Behav. 64:823–828.

Bolund, E., H. Schielzeth, and W. Forstmeier. 2007. Intrasexual competition in zebra finches, the role of beak colour and body size. Anim. Behav. 74:715–724.

Brodsky, L. M. 1988. Ornament size influences mating success in male rock ptarmigan. Anim. Behav. 36:662–667.

Buechner, V. L., M. A. Maier, S. Lichtenfeld, and A. J. Elliot. 2015. Emotion expression and color: their joint influence on perceived attractiveness and social position. Curr. Psychol. 34:422–433.

Burley, N. 1981. Sex-ratio manipulation and selection for attractiveness. Science 211:721–722.

———. 1985a. Leg-band color and mortality patterns in captive breeding populations of zebra finches. Auk 102:647–651.

———. 1985b. The organization of behavior and the evolution of sexually selected traits. Avian Monogamy 37:22–44.

———. 1986a. Sex-ratio manipulation in color-banded populations of zebra finches. Evolution 40:1191–1206.

———. 1986b. Sexual selection for aesthetic traits in species with biparental care. Am. Nat. 127:415–445.

———. 1988. The differential-allocation hypothesis: an experimental test. Am. Nat. 132:611–628.

Burley, N., G. Krantzberg, and P. Radman. 1982. Influence of color-banding on the conspecific preferences of zebra finches. Anim. Behav. 30:444–455.

Burley, N. T., D. A. Enstrom, and L. Chitwood. 1994. Extra-pair relations in zebra finches—differential male success results from female tactics. Anim. Behav. 48:1031–1041.

Burley, N. T., P. G. Parker, and K. Lundy. 1996. Sexual selection and extra-pair fertilization in a socially monogamous passerine, the zebra finch (Taeniopygia guttata). Behav. Ecol. 7:218–226.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14:444–444.

Cohen, J. 1988. Statistical power analysis for the behavioral science. Vol. 2. Hillsdale, NY: Lawrence Erlbaum Associates.

Collaboration, O. S. 2015. Estimating the reproducibility of psychological science. Science 349:aac4716.

Collins, S. A., and C. ten Cate. 1996. Does beak colour affect female preference in zebra finches? Anim. Behav. 52:105–112.

Cresswell, W., J. Lind, J. L. Quinn, J. Minderman, and D. P. Whitfield. 2007. Ringing or colour-banding does not increase predation mortality in redshanks *Tringa totanus*. J. Avian Biol. 38:309–316.

Cristol, D. A., C. S. Chiu, S. M. Peckham, and J. F. Stoll. 1992. Color bands do not affect dominance status in captive flocks of wintering dark-eyed juncos. Condor 94:537–539.

Cuthill, I. C., S. Hunt, C. Cleary, and C. Clark. 1997. Colour bands, dominance, and body mass regulation in male zebra finches (*Taeniopygia guttata*). Proc. R. Soc. B Biol. Sci. 264:1093–1099.

David, M., and F. Cézilly. 2011. Personality may confound common measures of mate-choice. Plos One 6:e24778.

Dechaume-Moncharmont, F. X., T. Brom, and F. Cezilly. 2016. Opportunity costs resulting from scramble competition within the choosy sex severely impair mate choosiness. Anim. Behav. 114:249–260.

Elliot, A. J., and M. A. Maier. 2012. 2 Color-in-context theory. Adv. Exp. Soc. Psychol. 45:61.

———. 2013. The red-attractiveness effect, applying the Ioannidis and Trikalinos (2007b) test, and the broader scientific context: a reply to Francis (2013). J. Exp. Psychol. Gen. 142:297–300.

Elliot, A. J., and D. Niesta. 2008. Romantic red: red enhances men's attraction to women. J. Pers. Soc. Psychol. 95:1150–1164.

Elliot, A. J., and A. D. Pazda. 2012. Dressed for sex: red as a female sexual signal in humans. Plos One 7:e34607.

Elliot, A. J., D. N. Kayser, T. Greitemeyer, S. Lichtenfeld, R. H. Gramzow, M. A. Maier, and H. J. Liu. 2010. Red, rank, and romance in women viewing men. J. Exp. Psychol. Gen. 139:399–417.

Elliot, A. J., T. Greitemeyer, and A. D. Pazda. 2013a. Women's use of red clothing as a sexual signal in intersexual interaction. J. Exp. Soc. Psychol. 49:599–602.

Elliot, A. J., J. L. Tracy, A. D. Pazda, and A. T. Beall. 2013b. Red enhances women's attractiveness to men: first evidence suggesting universality. J. Exp. Soc. Psychol. 49:165–168.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. Statistical power analyses using G* power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41:1149–1160.

Ferguson, C. J., and M. Heene. 2012. A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. Perspect. Psychol. Sci. 7:555–561.

Forstmeier, W. 2017. Preregister now for an upgrade to behavioral ecology 2.0: a comment on Ihle et al. Behav. Ecol. 28:358–359.

Forstmeier, W., and T. R. Birkhead. 2004. Repeatability of mate choice in the zebra finch: consistency within and between females. Anim. Behav. 68:1017–1028.

Forstmeier, W., K. Martin, E. Bolund, H. Schielzeth, and B. Kempenaers. 2011. Female extrapair mating behavior can evolve via indirect selection on males. Proc. Natl. Acad. Sci. USA 108:10608–10613.

Forstmeier, W., E. J. Wagenmakers, and T. H. Parker. 2017. Detecting and avoiding likely false-positive findings—a practical guide. Biol. Rev. 92:1941–1968.

Francis, G. 2013. Publication bias in Red, rank, and romance in women viewing men by Elliot et al. J. Exp. Psychol. Gen. 142:292–296.

Franco, A., N. Malhotra, and G. Simonovits. 2014. Publication bias in the social sciences: unlocking the file drawer. Science 345:1502–1505.

Freedman, L. P., I. M. Cockburn, and T. S. Simcoe. 2015. The economics of reproducibility in preclinical research. Plos Biol. 13:e1002165.

Gil, D., J. Graves, N. Hazon, and A. Wells. 1999. Male attractiveness and differential testosterone investment in zebra finch eggs. Science 286:126–128.

Greenwald, A. G. 1975. Consequences of prejudice against null hypothesis. Psychol. Bull. 82:1–19.

Griffith, S. C., and K. L. Buchanan. 2010. The zebra finch: the ultimate Australian supermodel. EMU:Austral Ornithol. 110:V–xii.

Hannon, S. J., and P. Eason. 1995. Color bands, combs and coverable badges in willow ptarmigan. Anim. Behav. 49:53–62.

Hesslinger, V. M., L. Goldbach, and C. C. Carbon. 2015. Men in red: a reexamination of the red-attractiveness effect. Psychon. B Rev. 22:1142–1148.

Holman, C., S. K. Piper, U. Grittner, A. A. Diamantaras, J. Kimmelman, B. Siegerink, and U. Dirnagl. 2016. Where have all the rodents gone? The effects of attrition in experimental research on cancer and stroke. PLoS Biol. 14:e1002331.

Horton, R. 2015. Offline: what is medicine's 5 sigma? Lancet 385:1380–1380.

Hunt, S., I. C. Cuthill, J. P. Swaddle, and A. T. D. Bennett. 1997. Ultraviolet vision and band-colour preferences in female zebra finches, *Taeniopygia guttata*. Anim. Behav. 54:1383–1392.

Ihle, M., B. Kempenaers, and W. Forstmeier. 2015. Fitness benefits of mate choice for compatibility in a socially monogamous species. PLoS Biol. 13:e1002248.

Ihle, M., I. S. Winney, A. Krystalli, and M. Croucher. 2017. Striving for transparent and credible research: practical guidelines for behavioral ecologists. Behav Ecol. 28:348–354.

Jennions, M. D. 1998. The effect of leg band symmetry on female-male association in zebra finches. Anim. Behav. 55:61–67.

Jennions, M. D., and A. P. Moller. 2002. Publication bias in ecology and evolution: an empirical assessment using the "trim and fill" method. Biol. Rev. 77:211–222.

Jerónimo, S., M. Khadraoui, D. Wang, K. Martin, J. A. Lesku, K. A. Robert, E. Schlicht, W. Forstmeier, and B. Kempenaers. 2018. Plumage color manipulation has no effect on social dominance or fitness in zebra finches. Behav Ecol.:arx195.

Johnsen, A., P. Fiske, T. Amundsen, J. T. Lifjeld, and P. A. Rohde. 2000. Colour bands, mate choice and paternity in the bluethroat. Anim. Behav. 59:111–119.

Johnsen, A., J. T. Lifjeld, and P. A. Rohde. 1997. Coloured leg bands affect male mate-guarding behaviour in the bluethroat. Anim. Behav. 54:121–130.

Kayser, D. N., A. J. Elliot, and R. Feltman. 2010. Red and romantic behavior in men viewing women. Eur. J. Soc. Psychol. 40:901–908.

Kinghorn, B., and A. Kinghorn. 2010. Pedigree viewer 6.5. University of New England, Armidale, Australia.

Knief, U., H. Schielzeth, N. Backstrom, G. Hemmrich-Stanisak, M. Wittig, A. Franke, S. C. Griffith, H. Ellegren, B. Kempenaers, and W. Forstmeier. 2017. Association mapping of morphological traits in wild and captive zebra finches: reliable within, but not between populations. Mol. Ecol. 26:1285–1305.

Kousta, S., C. Ferguson, E. Ganley, and P. B. Staff. 2016. Meta-research: broadening the scope. PLoS Biol 14:e1002334.

Lehmann, G. K., and R. J. Calin-Jageman. 2017. Is red really romantic? Soc. Psychol. 48:174–183.

Lenhard, W., and A. Lenhard. 2016. Calculation of effect sizes. Dettelbach (Germany): Psychometrica. https://doi.org/10.13140/RG.2.1.3478.4245.

McGraw, K. J., G. E. Hill, and A. J. Keyser. 1999. Ultraviolet reflectance of colored plastic leg bands. J. Field Ornithol. 70:236–243.

Metz, K. J., and P. J. Weatherhead. 1991. Color bands function as secondary sexual traits in male red-winged blackbirds. Behav. Ecol. Sociobiol. 28:23–27.

Mora, A. R., and W. Forstmeier. 2014. The importance of validating experimental setups: lessons from studies of food choice copying in zebra finches. Ethology 120:913–922.

Nakagawa, S., and J. R. Waas. 2004. The effect of acoustic and visual priming stimuli on the reproductive behaviour of female zebra finches, *Taeniopygia guttata*. Acta Ethol. 7:43–49.

Nakagawa, S., R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M. Senior. 2015. Meta-analysis of variation: ecological and evolutionary applications and beyond. Methods Ecol. Evol. 6:143–152.

Pariser, E. C., M. M. Mariette, and S. C. Griffith. 2010. Artificial ornaments manipulate intrinsic male quality in wild-caught zebra finches (*Taeniopygia guttata*). Behav. Ecol. 21:264–269.

Parker, T. H., and S. Nakagawa. 2017. Practical models for publishing replications in behavioral ecology: a comment on Ihle et al. Behav. Ecol. 28:355–357.

Parker, T. H., W. Forstmeier, J. Koricheva, F. Fidler, J. D. Hadfield, Y. E. Chee, C. D. Kelly, J. Gurevitch, and S. Nakagawa. 2016. Transparency in ecology and evolution: real problems, real solutions. Trends Ecol. Evol. 31:711–719.

Pazda, A. D., A. J. Elliot, and T. Greitemeyer. 2012. Sexy red: perceived sexual receptivity mediates the red-attraction relation in men viewing woman. J. Exp. Soc. Psychol. 48:787–790.

Pazda, A. D., A. J. Elliot, and T. Greitemeyer. 2014a. Perceived sexual receptivity and fashionableness: separate paths linking red and black to perceived attractiveness. Color Res. Appl. 39:208–212.

Pazda, A. D., P. Prokop, and A. J. Elliot. 2014b. Red and romantic rivalry: viewing another woman in red increases perceptions of sexual receptivity, derogation, and intentions to mate-guard. Pers. Soc. Psychol. B 40:1260–1269.

Peperkoorn, L. S., S. C. Roberts, and T. V. Pollet. 2016. Revisiting the red effect on attractiveness and sexual receptivity: no effect of the color red on human mate preferences. Evol. Psychol. 14. https://doi.org/10.1177/1474704916673841.

Prinz, F., T. Schlange, and K. Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10:712–781.

Ratikainen, I. I., and H. Kokko. 2010. Differential allocation and compensation: who deserves the silver spoon? Behav. Ecol. 21:195–200.

R Core Team. 2015. R: a language and environment for statistical computing. Vienna, Austria; 2015.

Riebel, K. 2009. Song and female mate choice in zebra finches: a review. Adv. Stud. Behav. 40:197–238.

Roche, E. A., T. W. Arnold, J. H. Stucker, and F. J. Cuthbert. 2010. Colored plastic and metal leg bands do not affect survival of Piping Plover chicks. J Field Ornithol 81:317–324.

Schuett, W., and S. R. X. Dall. 2010. Appearance, "state," and behavior in male zebra finches, *Taeniopygia guttata*. J. Ethol. 28:273–286.

Schwarzer, G., and M. G. Schwarzer. 2017. Package "meta". Meta-analysis with R:2.1–4.

Seguin, A., and W. Forstmeier. 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.

Sheldon, B. C. 2000. Differential allocation: tests, mechanisms and implications. Trends Ecol. Evol. 15:397–402.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22:1359–1366.

Song, F. J., and S. Gilbody. 1998. Bias in meta-analysis detected by a simple, graphical test—increase in studies of publication bias coincided with increasing use of meta-analysis. Brit. Med. J. 316:471–471.

Song, Z., Y. Liu, I. Booksmythe, and C. Ding. 2017. Effects of individual-based preferences for colour-banded mates on sex allocation in zebra finches. Behav. Ecol. 28:1228–1235.

Sossinka, R. 1980. Ovarian development in an opportunistic breeder, the zebra finch Poephila-Guttata-Castanotis. J. Exp. Zool. 211:225–230.

Spencer, K. A., and S. Verhulst. 2007. Delayed behavioral effects of postnatal exposure to corticosterone in the zebra finch (*Taeniopygia guttata*). Horm. Behav. 51:273–280.

Sterne, J. A. C., and M. Egger. 2001. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J. Clin. Epidemiol. 54:1046–1055.

Szucs, D., and J. P. A. Ioannidis. 2017. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15:e2000797.

Tang, J. L., and J. L. Y. Liu. 2000. Misleading funnel plot for detection of bias in meta-analysis. J. Clin. Epidemiol. 53:477–484.

Verner, J., D. Breese, and K. L. Purcell. 2000. Return rates of banded granivores in relation to band color and number of bands worn. J. Field Ornithol. 71:117–125.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36:1–48.

von Engelhardt, N. B. 2004. Proximate control of avian sex allocation: a study on zebra finches. Ph.D. thesis, University of Groningen, The Netherlands.

Wang, D., W. Forstmeier, and B. Kempenaers. 2017a. No mutual mate choice for quality in zebra finches: time to question a widely-held assumption. Evolution 71:2661–2676.

Wang, D., N. Kempenaers, B. Kempenaers, and W. Forstmeier. 2017b. Male zebra finches have limited ability to identify high-fecundity females. Behav. Ecol. 28:784–792.

Associate Editor: J. D. Hadfield
Handling Editor: M. A. F. Noor

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. Summary of publications ($n = 39$) [4-42] from studies in which male zebra finches were fitted with red versus green color bands.
Table S2. Mean relative fitness of female zebra finches with different color bands.
Table S3. Mean relative fitness of male zebra finches with different color bands.
Table S4. Linear mixed model explaining variation in reproductive success (square-root transformed number of independent offspring per breeding season) of 234 female zebra finches (excluding experiments 3, 4 and 6).
Table S5. Linear mixed model explaining variation in reproductive success (square-root transformed number of independent offspring sired per breeding season) of 234 male zebra finches (excluding experiments 3, 4 and 6).
Table S6. Linear mixed model of female fitness (version 1) using the "attractiveness rank" as a fixed effect (defined for six colors only).
Table S7. Linear mixed model of female fitness (version 2) using the "attractiveness rank" as a fixed effect (defined for six colors only).
Table S8. Linear mixed model of male fitness (version 1) using the "attractiveness rank" as a fixed effect (defined for six colors only).
Table S9. Linear mixed model of male fitness (version 2) using the "attractiveness rank" as a fixed effect (defined for six colors only).