Mate choice and the evolution of female promiscuity in a socially monogamous species

Dissertation der
Fakultät für Biologie der
Ludwig-Maximilians-Universität München

durchgeführt am
Max-Planck-Institut für Ornithologie
Seewiesen

vorgelegt von
Daiping Wang
August 2018

Mate choice and the evolution of female promiscuity in a socially monogamous species

Dissertation der
Fakultät für Biologie der
Ludwig-Maximilians-Universität München

durchgeführt am
Max-Planck-Institut für Ornithologie
Seewiesen

vorgelegt von
Daiping Wang
August 2018

Erstgutachter: Prof. Dr. **Bart Kempenaers**

Zweitgutachter: Prof. Dr. **Jochen Wolf**

Eingereicht am: 7.8.2018
Tag der mündlichen Prüfung: 5.2.2019

Diese Dissertation wurde unter der Leitung von Dr. **Wolfgang Forstmeier** angefertigt.

Contents

General introduction

Sexual selection and mate choice

In his book *On the Origin of Species by Means of Natural Selection,* Charles Darwin first proposed the concept of sexual selection. In the book (p 87-87, CHAP. IV., Darwin 1859) he wrote:

> "And this leads me to say a few words on what I call Sexual Selection. This depends, not on a struggle for existence, but on a struggle between the males for possession of the females; the result is not death to the unsuccessful competitor, but few or no offspring."

However, the more clear definition and detailed description of sexual selection came later in his other famous book *The Descent of Man, and Selection in Relation to Sex*. For instance, in that book (p 254-255, Part I. Darwin 1871) he said:

> ''We are, however, here concerned only with that kind of selection, which I have called sexual selection. This depends on the advantage which certain individuals have over other individuals of the same sex and species, in exclusive relation to reproduction.''

Further, Darwin suggested that (p 398 GENERAL SUMMARY Part II. Darwin 1871):

> "The sexual struggle is of two kinds; in the one it is between the individuals of the same sex, generally the male sex, in order to drive away or kill their rivals, the females remaining passive; whilst in the other, the struggle is likewise between the individuals of the same sex, in order to excite or charm those of the opposite sex, generally the females, which no longer remain passive, but select the more agreeable partners."

Progress after Darwin

In the more than 150 years after Darwin's propositions, much progress has achieved in developing and supporting the core parts of sexual selection, either theoretically by modeling or

practically with empirical examples (Andersson and Simmons 2006; Jones and Ratterman 2009). The first core part of sexual selection, that is 'male-male combat' (intra-sexual selection), seems easy to understand because of many solid examples such as male kudus *Tragelaphus strepsiceros* enormous horns (Davies et al. 2012) and much bigger size relative to females (3 to 7.5 times as heavy) of the male northern elephant seal *Mirounga angustirostris* (Le Boeuf and Reiter 1988). However, female choice, perhaps more specifically, why females are choosy, the other core part of Darwin's sexual selection, has intrigued many behavioral ecologists for decades (Andersson 1994). Under a framework of adaption which focuses on benefits the female will gain from choosiness, models have been developed for its evolutionary explanation. Each kind of model is supported by some empirical examples (Jones and Ratterman 2009; Davies et al. 2012).

a. direct-benefits models

The direct-benefits models suppose that females (or males in sex-role-reversed species) could benefit directly from their chosen mates, through better parental care, a nuptial gift, or territory defence. For instance, female North American bullfrogs *Rana catesbeiana* (Howard 1978a, b) choose males that have good territories and lay their eggs in those territories, which can increase the survival of eggs. Males of the bushcricket *Ephippiger ephippiger* (Gwynne 1984) and hanging fly *Hylobittacus apicalis* (Dussourd et al. 1991) provide a nuptial gift to their mates that the female can eat during or after copulation. Evolution of choice for direct benefits is conceptually simple because the advantage resulting from choosing is obvious. Nevertheless, one point needed to be kept in mind for these direct-benefits models is that male-male competition often goes hand in hand with female choice in the process of providing direct benefits. For example, male northern elephant seals which are bigger relative to females have more chances to win against other males. At the same time, bigger male seals also provide better protection to their harems.

b. indirect-benefits models

In earlier years, the Fisherian Models (Fisher 1930; Kirkpatrick 1982; Mead and Arnold 2004) were one kind of representative indirect-benefits model explaining the evolution of female choice of one specific male trait. Assuming female preference for a male trait (no matter if this male trait is a reliable quality indicator or just attractive to females), as soon as this preference of the male trait leads to genetic benefits to females, female mate choice will result in a genetic correlation between the female preference and the male trait. This genetic correlation will develop into positive feedback between female preference and the male trait until conflicts arise between sexual selection and natural selection. Evidence for this model came from a lekking sandfly *Lutzomyia longipalpis* which showed generally attractive males fathered sons

who were then chosen when they in turn formed leks (Jones et al. 1998). Other supportive studies showed there is a positive genetic correlation between preference and a male trait: stalk-eyed fly *Cyrtodiopsis dalmanni*, (Wilkinson and Reillo 1994); guppy *Poecilia reticulate* (Houde and Endler 1990; Brooks 2000) and three-spined stickleback *Gasterosteus aculeatus* (Milinski and Bakker 1990).

Perhaps the most famous indirect-benefits model is the 'good genes' with a more adaptive way of thinking (Zahavi 1975, 1977; Hamilton and Zuk 1982). The assumption of this model requires the male trait to be a reliable quality indicator (e.g. a costly ornament) or indicates good genes (e.g. an ornament genetically correlated to viability traits). Female choice evolves because females who chose the male trait (e.g. more elaborately ornamented male) could produce offspring with higher viability or that will be in good condition as adults. This model enjoys the most empirical support by showing the phenotypic correlation between the focal male trait and male reproductive traits. One famous example is the extraordinary tail of the male Indian peafowl (*Pavo cristatus*) which signals a male's genetic quality (Petrie et al. 1991; Petrie 1994). However, even though this is such a famous example, in another study of this species, the authors did not find that females preferred males with more elaborate tails (Takahashi et al. 2008).

c. other models

Besides direct-benefits and indirect-benefits models, several other models have been proposed for explanation of female choice. For instance, a class of models focused on the genetic compatibility between female and male mates. Polymorphic genes of the major histocompatibility complex (MHC) are regarded as essential genes for individual fitness under conditions of natural and sexual selection (Milinski 2006). Studies showed that, with reference to their own MHC profile, female sticklebacks preferred to mate with a male sharing an intermediate MHC diversity to get an optimal complement. Therefore, this could provide resistance against parasites, which could be revealed by the expression of costly secondary sexual characters (Eizaguirre et al. 2009). Another class of models suggested that males evolve sexually selected traits because of the preexisting inclinations of female sensory systems (the sensory exploitation model, Endler and Basolo 1998; Ryan 1998). This sensory bias inherent to the choosing females might result from random drift or some other evolutionary drive (e.g. natural selection, Fuller et al. 2005).

Where do we go next?

In short, given the amount of theoretic and empirical effort, it has seemed rather fruitful in the field of sexual selection and mate choice since Darwin. Thus, summarization (see above) of these theoretic models and practical evidence in the past 150 years provides us with rather promisingly future directions.

a. is the pattern of mate choice in monogamous species the same?

The famous examples explaining why females should choose the 'more agreeable partners' (Darwin 1871) in the mating pool came from polygynous species (or polyandrous in sex-role-reversed), for instance, the much bigger size of the male northern elephant seal relative to female (Le Boeuf and Reiter 1988), extraordinarily long tail of the male widowbird (Andersson 1982) and spectacular displaying of the male peacock (Petrie et al. 1991). All these species are 'lekking' species with little parental contribution from the male partner. In these cases, the benefits of female choosiness (e.g. good-gene or sexy-son benefits) seem obvious. Furthermore, there are little costs of being choosy because it is easy to mate the preferred male with little female-female competition. Consequently, hypotheses proposing that females are always choosy and will prefer the highest-quality male seemly dominated the field of sexual selection and thus have spread to mate choice of monogamous species.

However, in socially monogamous species with bi-parental care (expected to favor choosiness, Kokko and Johnstone 2002), the situation is more complicated and subtle. First, males in these species contribute a relatively equal amount of parental care compared to the females. This implies males might be choosy as well (see **Chapter 1**). Second, if all females are choosy and have consensus on the highest-quality male in the population, this means intensive female-female competition to pair with the best male. Logically following the rational, if only the high quality females could pair with those high quality males (assortative mating for quality), the rest of females will be left as unpaired or incompatible with their partners if they paired with the remaining males in the population. If this would be the case, at the level of population, the mechanism might not be evolutionary stable. Given that most individuals of monogamous species in the field formed breeding pairs rather unpaired, other more stable and subtle mechanisms of mate choice might drive the mating pattern (see **Chapter 2**).

b. is the male trait really a reliable quality indicator?

A large body of mate choice literature has focused on documenting the extent to which ornaments or displays can function as honest signals of intrinsic quality or current condition (Hamilton and Zuk 1982; Nakagawa et al. 2007; Catchpole and Slater 2008; Dunn et al. 2010), and numerous studies have described directional mating preferences for such quality indicators

(Andersson 1982; Welch et al. 1998; Reid et al. 2004; Pincemy et al. 2009; Doutrelant et al. 2012; Wells et al. 2015). It is therefore tempting to assume that directional mate choice for quality indicators will be ubiquitous in the animal kingdom. Yet this assumption can be challenged for several reasons (see **Chapter 2**). Logically, if the male trait is an honest signal of quality, then the trait values could explain a large proportion of male fitness in population. In this respect, studies focusing on to which extent the trait could explain the true variance in quality (e.g. fitness) seem more valuable before checking the female choice of that trait. In fact, the condition-dependence of quality indicators is often limited (Cotton et al. 2004; Bolund et al. 2010; **Chapter 1**) and requires more concern.

c. are the text book examples of sexual selection reproducible?

Many fields of science - including behavioral ecology – are currently experiencing a heated debate about the extent to which publication bias against null-findings results in a misrepresentative scientific literature. Specifically, in studies of mate choice, for each studied species, numerous potential quality indicators can be measured and preferences or choice outcomes can be quantified in many different ways. In empirical studies, this often leads to a considerable problem of multiple testing in combination with the risk of selective reporting of positive results (Forstmeier et al. 2016). This makes it difficult to judge how often the null hypothesis of no directional preference for quality indicators might actually be true. Take, for instance, the example of the extraordinarily long tail of the peacock. In an English study, Marion Petrie (1991, 1994) found that the tail display of this species can predict a male's mating success, and it is a reliable indicator of genetic quality. However, another study in Japan, did not find evidence that females choose males with more elaborate tails (Takahashi et al. 2008). An explanation for the discrepancy between these two studies is that female choice varies in different contexts. However, in evolutionary biology, biological conclusions should be formulated with caution under the condition of context-dependence (**Chapter 3, Chapter 4**).

Study species and thesis outline

I used zebra finches (*Taeniopygia guttata*, Figure 1) to study sexual selection during my PhD. Zebra finches are one of the most intensely studied organisms regarding mate choice (reviewed by Collins and ten Cate 1996; Adkins-Regan 1998; ten Cate and Vos 1999; Riebel 2003; Griffith and Buchanan 2010; Hauber et al. 2010; Adkins-Regan 2011). Their predominant mating system is lifetime monogamy with both sexes investing about equally in parental care (Zann 1996). This high investment of both sexes is expected to favor choosiness in both males and females when searching for a (lifetime) partner (Kokko and Johnstone 2002). The species is abundant, breeds

in dense colonies, and forms large flocks in the non-breeding period, where new pair bonds can form long before reproduction (Zann 1996). This means that encounter rates of potential mates are presumably high, and hence the cost of being choosy during the period of pair formation should be low and should not hamper the evolution of choosiness (Johnstone 1997; Kokko and Johnstone 2002). Inspired by the ideas mentioned above (see details in section of 'where do we go next'), there are three topic words throughout my thesis outlining five chapters.

Figure 1: Two male and a female (middle) zebra finches. Photo from Wolfgang Forstmeier

The first word is 'role'. Underlying this word, I systematically investigate mate choice in this species from different sexual perspectives to assess the role of each sex during the choice process. Specifically, in **Chapter 1**, I studied male mate choice for a trait of female fitness (female fecundity). Further, in species such as zebra finches where both males and females invest substantially in parental care, we expect both sexes to be choosy. Under the assumption of preferences for high-quality individuals, mutual mate choice will then result in assortative mating by quality. This led to **Chapter 2** which investigated mutual mate choice in zebra finches.

The second topic word is 'scale' with which I aim to study mate choice in monogamous species at different scales. In **Chapter 1** and **2**, I studied mate choice within a captive zebra finch population. In **Chapter 3**, with data from seven different populations, I tested the reliability and

generality of a textbook example of mate choice in this species. **In Chapter 4**, we did a meta-analysis of assortative mating in birds which included published data from 133 species and unpublished data from nine long-term-study species.

The third topic word is 'replication' meaning to test reproducibility of key findings in this model species. Previous experimental work on zebra finches has shown that males preferred females whose fecundity had been boosted by a high-protein diet (Monaghan et al. 1996; Jones et al. 2001). However, it remained unclear whether the demonstrated ability to identify protein-supplemented females would extend to an ability to assess non experimental variation in female fecundity that exists under a standardized diet. Thus, **Chapter 1** addressed these issues, by quantifying the extent to which male zebra finches are able to perceive normal variation in female fecundity using a two-way choice paradigm. In **Chapter 2**, given that a study reported significant assortative mating for a putative quality indicator in this species (Holveck and Riebel 2010), I used principal component analysis (PCA) to summarize 10 quality-related traits measured in male and six quality-related traits in female zebra finches into a single quality score to test for assortative mating for quality. In **Chapter 3**, using a couple of different populations from different labs, I replicated a text-book example of mate choice in this species. In **Chapter 5**, I used a better experimental design with more sophisticated data to further explore and verify a previous finding in our group which showed the genetic constraints of female promiscuity (Forstmeier et al. 2011).

References

Adkins-Regan, E. 1998. Hormonal mechanisms of mate choice. Am Zool 38:166-178.

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

Andersson, M. 1982. Female Choice Selects for Extreme Tail Length in a Widowbird. Nature 299:818-820.

Andersson, M. and L. W. Simmons. 2006. Sexual selection and mate choice. Trends Ecol Evol 21:296-302.

Andersson, M. B. 1994. Sexual selection. Princeton University Press, Princeton, N.J.

Bolund, E., H. Schielzeth, and W. Forstmeier. 2010. No heightened condition dependence of zebra finch ornaments - a quantitative genetic approach. J Evolution Biol 23:586-597.

Brooks, R. 2000. Negative genetic correlation between male sexual attractiveness and survival. Nature 406:67-70.

Catchpole, C. and P. Slater. 2008. Bird song: biological themes and variations. Cambridge university press. 2nd ed. .

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

Cotton, S., K. Fowler, and A. Pomiankowski. 2004. Do sexual ornaments demonstrate heightened condition-dependent expression as predicted by the handicap hypothesis? P Roy Soc B-Biol Sci 271:771-783.

Darwin, C. 1859. On the origin of species by means of natural selection. (Murray, London).

Darwin, C. 1871. The descent of man, and selection in relation to sex. (Murray, London). . Murray.

Davies, N. B., J. R. Krebs, and S. A. West. 2012. An introduction to behavioural ecology. John Wiley & Sons.

Doutrelant, C., A. Gregoire, A. Midamegbe, M. Lambrechts, and P. Perret. 2012. Female plumage coloration is sensitive to the cost of reproduction. An experiment in blue tits. J Anim Ecol 81:87-96.

Dunn, P. O., J. C. Garvin, L. A. Whittingham, C. R. Freeman-Gallant, and D. Hasselquist. 2010. Carotenoid and melanin-based ornaments signal similar aspects of male quality in two populations of the common yellowthroat. Funct Ecol 24:149-158.

Dussourd, D. E., C. A. Harvis, J. Meinwald, and T. Eisner. 1991. Pheromonal Advertisement of a Nuptial Gift by a Male Moth (Utetheisa-Ornatrix). P Natl Acad Sci USA 88:9224-9227.

Eizaguirre, C., S. E. Yeates, T. L. Lenz, M. Kalbe, and M. Milinski. 2009. MHC ‐ based mate choice combines good genes and maintenance of MHC polymorphism. Mol Ecol 18:3316-3329.

Endler, J. A. and A. L. Basolo. 1998. Sensory ecology, receiver biases and sexual selection. Trends Ecol Evol 13:415-420.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection

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

Forstmeier, W., E. J. Wagenmakers, and T. H. Parker. 2016. Detecting and avoiding likely false ‐ positive findings – a practical guide. Biological Reviews.

Fuller, R. C., D. Houle, and J. Travis. 2005. Sensory bias as an explanation for the evolution of mate preferences. Am Nat 166:437-446.

Griffith, S. C. and K. L. Buchanan. 2010. The Zebra Finch: the ultimate Australian supermodel. Emu 110:V-xii.

Gwynne, D. T. 1984. Courtship Feeding Increases Female Reproductive Success in Bush-Crickets. Nature 307:361-363.

Hamilton, W. D. and M. Zuk. 1982. Heritable True Fitness and Bright Birds - a Role for Parasites. Science 218:384-387.

Hauber, M. E., D. L. M. Campbell, and S. M. N. Woolley. 2010. The functional role and female perception of male song in Zebra Finches. Emu 110:209-218.

Holveck, M. J. and K. Riebel. 2010. Low-quality females prefer low-quality males when choosing a mate. P Roy Soc B-Biol Sci 277:153-160.

Houde, A. E. and J. A. Endler. 1990. Correlated Evolution of Female Mating Preferences and Male Color Patterns in the Guppy Poecilia-Reticulata. Science 248:1405-1408.

Howard, R. D. 1978a. Evolution of Mating Strategies in Bullfrogs, Rana-Catesbeiana. Evolution 32:850-871.

Howard, R. D. 1978b. Influence of Male-Defended Oviposition Sites on Early Embryo Mortality in Bullfrogs. Ecology 59:789-798.

Johnstone, R. A. 1997. The tactics of mutual mate choice and competitive search. Behav Ecol Sociobiol 40:51-59.

Jones, A. G. and N. L. Ratterman. 2009. Mate choice and sexual selection: What have we learned since Darwin? P Natl Acad Sci USA 106:10001-10008.

Jones, K. M., P. Monaghan, and R. G. Nager. 2001. Male mate choice and female fecundity in zebra finches. Anim Behav 62:1021-1026.

Jones, T. M., R. J. Quinnell, and A. Balmford. 1998. Fisherian flies: benefits of female choice in a lekking sandfly. P Roy Soc B-Biol Sci 265:1651-1657.

Kirkpatrick, M. 1982. Sexual Selection and the Evolution of Female Choice. Evolution 36:1-12.

Kokko, H. and R. A. Johnstone. 2002. Why is mutual mate choice not the norm? Operational sex ratios, sex roles and the evolution of sexually dimorphic and monomorphic signalling. Philos T Roy Soc B 357:319-330.

Le Boeuf, B. and J. Reiter. 1988. Lifetime reproductive success in northern elephant seals. Reproductive success. University of Chicago Press, Chicago:344-362.

Mead, L. S. and S. J. Arnold. 2004. Quantitative genetic models of sexual selection. Trends Ecol Evol 19:264-271.

Milinski, M. 2006. The major histocompatibility complex, sexual selection, and mate choice. Annu Rev Ecol Evol S 37:159-186.

Milinski, M. and T. C. M. Bakker. 1990. Female Sticklebacks Use Male Coloration in Mate Choice and Hence Avoid Parasitized Males. Nature 344:330-333.

Monaghan, P., N. B. Metcalfe, and D. C. Houston. 1996. Male finches selectively pair with fecund females. P Roy Soc B-Biol Sci 263:1183-1186.

Nakagawa, S., N. Ockendon, D. O. S. Gillespie, B. J. Hatchwell, and T. Burke. 2007. Assessing the function of house sparrows' bib size using a flexible meta-analysis method. Behav Ecol 18:831-840.

Petrie, M. 1994. Improved Growth and Survival of Offspring of Peacocks with More Elaborate Trains. Nature 371:598-599.

Petrie, M., T. Halliday, and C. Sanders. 1991. Peahens Prefer Peacocks with Elaborate Trains. Anim Behav 41:323-331.

Pincemy, G., F. S. Dobson, and P. Jouventin. 2009. Experiments on colour ornaments and mate choice in king penguins. Anim Behav 78:1247-1253.

Reid, J. M., P. Arcese, A. L. E. V. Cassidy, S. M. Hiebert, J. N. M. Smith, P. K. Stoddard, A. B. Marr, and L. F. Keller. 2004. Song repertoire size predicts initial mating success in male song sparrows, Melospiza melodia. Anim Behav 68:1055-1063.

Riebel, K. 2003. Developmental influences on auditory perception in female zebra finches - is there a sensitive phase for song preference learning? Anim Biol 53:73-87.

Ryan, M. J. 1998. Sexual selection, receiver biases, and the evolution of sex differences. Science 281:1999-2003.

Takahashi, M., H. Arita, M. Hiraiwa-Hasegawa, and T. Hasegawa. 2008. Peahens do not prefer peacocks with more elaborate trains. Anim Behav 75:1209-1219.

ten Cate, C. and D. R. Vos. 1999. Sexual imprinting and evolutionary processes in birds: A reassessment. Advances in the Study of Behavior, Vol. 28 28:1-31.

Welch, A. M., R. D. Semlitsch, and H. C. Gerhardt. 1998. Call duration as an indicator of genetic quality in male gray tree frogs. Science 280:1928-1930.

Wells, S. J., W. H. Ji, J. Dale, B. Jones, and D. Gleeson. 2015. Male size predicts extrapair paternity in a socially monogamous bird with extreme sexual size dimorphism. Behav Ecol 26:200-206.

Wilkinson, G. S. and P. R. Reillo. 1994. Female Choice Response to Artificial Selection on an Exaggerated Male Trait in a Stalk-Eyed Fly. P Roy Soc B-Biol Sci 255:1-6.

Zahavi, A. 1975. Mate Selection - Selection for a Handicap. J Theor Biol 53:205-214.

Zahavi, A. 1977. Cost of Honesty - (Further Remarks on Handicap Principle). J Theor Biol 67:603-605.

Zann, R. A. 1996. The zebra finch : a synthesis of field and laboratory studies. Oxford University Press, Oxford ; New York.

Chapter 1: Male zebra fiches have limited ability to identify high-fecundity females

Short title: Male mate choice in zebra finches

Abstract: In species with bi-parental care and lifetime monogamy, the fecundity of a male's partner can be a major component of his fitness, but it is unclear whether males can assess female fecundity before breeding. We carried out an experiment in which we measured variation in female fecundity (repeatability 39%, 213 females) in a captive zebra finch population, and tested whether males preferred unfamiliar females of high fecundity (approximately top 10% of the population; 30 eggs laid on average) over those of low fecundity (bottom 10%; 6 eggs). We first tested whether naïve human observers could identify the high-fecundity female when confronted with duos of high and low fecundity. Humans guessed correctly in 58% of the cases (95% CI 50%-66%) indicating that differences in female condition were not highly obvious to humans. Zebra finch males preferred the high-fecundity female in 59% of choice tests that lasted 20 min (CI 52%-66%). When extending such choice tests over several days, male "success" in associating with the high-fecundity female was still modest (61% correct choices, CI 44%-76%). Overall, male zebra finches seem to have only limited abilities to identify the better mate when faced with a choice between extremes in terms of female fecundity. We found no male preference for heavier females. We speculate that such a preference may not have evolved because, in contrast to many ectothermic species, predicting fecundity from female weight is not sufficiently accurate (r^2 = 0.04) for the benefits to outweigh the costs of increased male-male competition for heavy females.

Published as: Daiping Wang, Nele Kempenaers, Bart Kempenaers, Wolfgang Forstmeier 2017: Male zebra finches have limited ability to identify high-fecundity females. Behav Ecol 28(3):784-792.

Behavioral Ecology

The official journal of the
ISBE
International Society for Behavioral Ecology

Behavioral Ecology (2017), 28(3), 784–792. doi:10.1093/beheco/arx037

Original Article

Male zebra finches have limited ability to identify high-fecundity females

Daiping Wang,[a] Nele Kempenaers,[a] Bart Kempenaers,[a] and Wolfgang Forstmeier[a]
[a]Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Str. 7, 82319 Seewiesen, Germany

Received 25 August 2016; revised 15 December 2016; editorial decision 2 February 2017; accepted 10 March 2017; Advance Access publication 16 March 2017.

In species with biparental care and lifetime monogamy, the fecundity of a male's partner can be a major component of his fitness but it is unclear whether males can assess female fecundity before breeding. We carried out an experiment in which we measured variation in female fecundity (repeatability 39%, 213 females) in a captive zebra finch population and tested whether males preferred unfamiliar females of high fecundity (approximately top 10% of the population; 30 eggs laid on average) over those of low fecundity (bottom 10%; 6 eggs). We first tested whether naïve human observers could identify the high-fecundity female when confronted with duos of high and low fecundity. Humans guessed correctly in 58% of the cases (95% confidence interval [CI] 50–66%) indicating that differences in female condition were not highly obvious to humans. Zebra finch males preferred the high-fecundity female in 59% of choice tests that lasted 20 min (CI 52–66%). When extending such choice tests over several days, male "success" in associating with the high-fecundity female was still modest (61% correct choices, CI 44–76%). Overall, male zebra finches seem to have only limited abilities to identify the better mate when faced with a choice between extremes in terms of female fecundity. We found no male preference for heavier females. We speculate that such a preference may not have evolved because, in contrast to many ectothermic species, predicting fecundity from female weight is not sufficiently accurate ($r^2 = 0.04$) for the benefits to outweigh the costs of increased male–male competition for heavy females.

Key words: **body size, female fecundity, male mate choice, mate choice cues, ornaments, preferences, quality indicators, sexual selection.**

INTRODUCTION

Over the last 4 decades, there has been a lively interest in the study of mate choice (Andersson 1994; Andersson and Simmons 2006; Charmantier and Sheldon 2006). In general, the sex that makes the greater reproductive investment should be the choosier sex (Trivers 1972; Clutton-Brock and Parker 1992) and in most cases, this is the female (Andersson 1994). Indeed, female mate choice has been studied extensively, particularly regarding potential benefits, such as "good gene benefits" (Zahavi 1977; von Schantz et al. 1999), "sexy son benefits" (Houde and Endler 1990), and direct benefits in terms of ensuring male fertility (Sheldon 1994; Mautz et al. 2013) or parental care (Hoelzer 1989; Alonzo 2012).

In species where males invest substantially in parental care, males are also expected to be choosy (Andersson 1994; Smiseth and Amundsen 2000). A preference for females of high fecundity may translate into substantial fitness gains for males (Edward and Chapman 2012), particularly in lifetime monogamous species where males typically reproduce only with a single female (Monaghan et al. 1996; Jones et al. 2001). However, relatively few studies have addressed male mate choice (Jones and Hunter 1993; Torres and Velando 2005; Griggio et al. 2009; Edward and Chapman 2011).

Although the potential benefits from male choice for highly fecund females are relatively large, directional selection via male mate choice requires an indicator trait that reliably signals female fecundity. In many taxonomic groups, in particular in ectotherms, females vary substantially in body size (e.g., Willemsen and Hailey 1999; Koops et al. 2004; Long et al. 2009) and this variation is often tightly correlated with variation in female fecundity (e.g., Bonduriansky 2001; Koops et al. 2004). Accordingly, male mate preferences for larger females have been well documented in at least some ectothermic species including insects (Edward and Chapman 2012), fish (Cote and Hunte 1989; Pelabon et al. 2003), amphibians (Arntzen 1999), and reptiles (Swierk et al. 2013). Endotherms, in contrast, typically show less variation in body size of adult (reproductively active) females (e.g., Zedrosser et al. 2006) and body size is typically a poor predictor of female fecundity (Jensen et al. 2004). In such species, reliable cues to female fecundity might not exist or they may be less obvious (to the researcher).

Address correspondence to W. Forstmeier. E-mail: forstmeier@orn.mpg.de.

© The Author 2017. Published by Oxford University Press on behalf of the International Society for Behavioral Ecology. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Experimental work on lifetime monogamous zebra finches, *Taeniopygia guttata*, has shown that males preferred females whose fecundity had been boosted by a high-protein diet (Monaghan et al. 1996; Jones et al. 2001). However, it remained unclear how males were able to assess female fecundity. Protein-supplemented females may have sent out behavioral signals indicating an increased readiness to mate and breed or diet may have affected female body mass, which males might have perceived during female movements or other female visual or even olfactory traits. It also remained unclear whether the demonstrated ability to identify protein-supplemented females would extend to an ability to assess nonexperimental variation in female fecundity that exists under a standardized diet.

The main aim of our study was to address these issues, by quantifying the extent to which male zebra finches are able to perceive normal variation in female fecundity using a 2-way choice paradigm. To maximize our ability to detect any effect on male mate choice, we selected stimulus females for the choice tests that differed markedly in fecundity. Specifically, we selected from the top and the bottom 10% of the population distribution in fecundity. Our experiment made use of another study where fecundity had been measured twice under standardized conditions in 4 successive groups of 54 females. This allowed us to conduct 2 identical replicates of the choice experiment (with females selected from a pool of 108 individuals in each replicate), in order to examine the reliability of our findings (Amundsen 2000; Nakagawa and Parker 2015). Moreover, we presented duos of high- versus low-fecundity females to naive human observers asking them to guess which of the 2 females is of high fecundity. This was done to investigate whether the 2 types of females differed in any way that is obvious to humans (e.g., differences in plumage condition or signs of sickness).

MATERIAL AND METHODS

Population and assessment of female fecundity

Details about our study population of domesticated zebra finches and about how we assessed them for variation in female fecundity are presented in the Supplementary Material. In brief, females were given the opportunity to lay eggs over a 7-week period in communal breeding aviaries that allowed free mate choice (aviaries contained 6 males and 6 females). All eggs were collected for parentage assignment and replaced with plastic eggs. Clutches of plastic eggs were removed after 10 days of incubation to allow the female to lay the next clutch. This 7-week breeding period was repeated with a different set of potential partners, which allowed us to quantify the repeatability of female fecundity. Birds were observed daily to derive 2 parameters of pairing success: the number of days that a female was socially paired ("days paired") and the exclusivity of her partner showing such pair bonding behavior only with her ("female share"). Daily nest checks combined with behavioral observations allowed us to assign 95% of all eggs laid (3840 out of 4041) to social parents that attended the respective nest. Social assignment of eggs was the basis on which we selected females of low and high fecundity ("estimated fecundity"). "True fecundity" was only assessed after the choice experiments by parentage analysis using 15 microsatellite markers (see Supplementary Table S1). Female age at the start of the breeding experiment (range 269–939 days) was a significant predictor of true fecundity ($r = -0.14$, $n = 213$ females, $P = 0.044$). This decline in fecundity with age suggests that males might benefit from preferring young females. Hence, age was considered in the analysis of choice tests (see below).

Fecundity analysis and selection of stimulus females

Within each replicate, we assessed individual differences in "estimated fecundity" using a mixed-effect model, with the number of eggs laid per 7-week breeding round as the dependent variable, with female identity (ID) as a random effect, and controlling for the fixed effects "breeding round," "days paired," and "female share." We used the "best linear unbiased predictors" (BLUPs) of fecundity for all females that were still alive and not obviously sick (replicate 1: $n = 101$, replicate 2: $n = 94$) to select the top and bottom 10 females within each replicate. By selecting 10 high- and 10 low-fecundity females according to their BLUPs in each replicate we identified females that had laid the most and the fewest eggs after controlling for their social pairing situation. For low-fecundity females, the model hence allowed us to identify those that laid few eggs despite being paired, rather than those that failed to pair and laid few eggs because of that. By using this approach, we might have missed some low-fecundity females but their true fecundity would have been uncertain and these females might be behaviorally peculiar (in each round, there were about 5 such females who were often unpaired and laid fewer eggs than some of the paired females that we selected). For high-fecundity females, this approach of fitting "days paired" as a fixed effect did not affect which of the females were selected, because most females (and all high-fecundity ones) paired soon after starting the experiment.

Then we formed 10 duos of stimulus females within each of the 2 replicates to be used in all choice tests by randomly combining 1 high- and 1 low-fecundity female. The above mixed-effect models were based on "estimated fecundity" but the parentage analysis confirmed that we had correctly identified duos of females with large differences in true fecundity (see Results for details, Table 1).

Two-way choice tests: male and female behavior

Before the start of choice tests, we weighed all females (nearest 0.1 g) and measured beak coloration using spectrophotometry. Six main characteristics of the reflectance spectrum were summarized to a discriminant axis score that separates the sexes as described in (Bolund et al. 2007; Schielzeth et al. 2012), with high values referring to male-like red coloration and low values to female-like orange. Body mass and beak color are condition-dependent traits affected by early growth conditions and inbreeding (Bolund, Martin, et al. 2010; Bolund, Schielzeth, et al. 2010), with higher mass and redder beaks indicating better condition. We calculated the difference in body mass, beak color scores, and age between the females of a duo (high fecundity minus low fecundity) and assessed its explanatory value for male preference for high-fecundity females. We expected that males would prefer females with higher mass, redder beaks, and younger females.

For each of the 20 female duos, we randomly selected 6 test males (120 different males in total). These males had the same background experience as the females, that is, they had participated in the aviary breeding experiment used for the assessment of female fecundity. However, we ensured that the 6 test males in each group were unfamiliar with and not closely related to both females they were exposed to in the choice experiment. Given that there was no significant difference in inbreeding coefficient between high- and low-fecundity females ($F \pm$ SD of high-fecundity females:

Table 1

Fecundity information for selected females in terms of fecundity in each of 2 replicate experiments

Experiment	Female fecundity	Number of females	"Estimated fecundity" mean ± SD (range)	BLUPs mean ± SD (range)	"True fecundity" mean ± SD (range)	Number of females	Eggs in nest building mean ± SD (range)
Replicate 1	High	10	30.2 ± 0.9 (29–32)	6.8 ± 1.4 (5.6–9.2)	29.4 ± 1.2 (27–31)	8	3.3 ± 1.5 (0–4.7)
	Low	10	2.8 ± 3.9 (0–10)	−9.1 ± 2.6 (−13–−6.3)	5.2 ± 4.5 (0–15)	8	0.5 ± 0.8 (0–2.3)
Replicate 2	High	10	29.9 ± 4.1 (22–37)	5.0 ± 1.1 (3.6–7.6)	29.8 ± 3.7 (23–35)	10	2.1 ± 1.3 (0.3–4.3)
	Low	10	4.2 ± 4.2 (0–12)	−6.0 ± 2.2 (−9.5–−3.7)	6.2 ± 6.6 (0–18)	10	0.8 ± 1.2 (0–4)

The "estimated fecundity" refers to the total number of eggs assigned to individual females (before parentage analysis), whereas "true fecundity" refers to assignment after parentage analysis (including infertile eggs that are still only socially assigned, see Methods for details). Females had been selected according to their BLUPs from the models shown in Supplementary Table S3. Here, these BLUPs are shown multiplied by 2 in order to reflect expectations for the sum of both breeding rounds (expected total number of eggs relative to the population mean). Egg-laying patterns of all selected females in the experiment are shown in Supplementary Figure S3. The last column shows the average number of eggs that the females laid during the nest-building experiment per experimental test (averaged across 3 choice tests for each female).

0.09 ± 0.05, low fecundity: 0.11 ± 0.06, $t = -0.94$, df = 19, $P = 0.36$) and zebra finches have no ability to judge relatedness beyond familiarity (Ihle and Forstmeier 2013), we did not consider inbreeding and relatedness further in this study.

Within each replicate, all 2-way choice tests took place over the course of 3 weeks. For each female duo, the 3 weeks of testing were arranged as follows (with the 6 test males designated as A-F): 1 test of 20 min per day, 4 tests from Monday to Thursday in each week, encountering males A-B-A-B in week 1, C-D-C-D in week 2, and E-F-E-F in week 3. The test order of the 10 duos within each day was randomized. We tested each male with the same duo twice (2 days apart) to allow calculating the repeatability of male preference for a particular female within the duo. Each choice test was composed of 2 halves of 10 min, whereby females were swapped between cages at halftime allowing us to differentiate between male preference and male side-bias. We allocated females randomly to cages at the start of a choice test and the observer of male and female behavior was blind to the information on female fecundity.

In replicate 1, 2 low-fecundity females died during the course of testing (1 by accident just before test #4 of 12, 1 naturally just before test #9 of 12) leading to the cancellation of 13 choice tests, leaving 227 tests involving 114 males.

The 2-way choice chamber used is a classical mate choice set-up (Supplementary Figure S1A) where the choosing male can spend time in the neutral zone where food is provided or can approach 1 of the females at either end of the apparatus while remaining separated from the female by wire mesh. Mate preference was assessed by recording the amount of time that a male spent outside the neutral zone with each female, facing the female and being active, which typically included directed courtship song (not counted is time spent inactively or facing away from the female, following [Rutstein et al. 2007]). In order to test whether more active males or males that were more interested in assessing or courting females made better choices, we summed up the times that males spent with each of the 2 females over the 20 min test period ("choosing motivation") and used it as an explanatory variable. The response variable of interest was calculated as the relative time each male spent with the high-fecundity female (ranging from 0 to 1, expected mean under the null hypothesis = 0.5). In 6 out of 227 tests, the male did not leave the neutral zone (4 tests where the male was active in the neutral zone, 2 tests where the male was completely inactive) leaving 221 informative tests involving 113 males.

During each choice test, the observer (D.W.) also recorded female responsiveness to the male, ranging from 0 (no signs of interest) to 1 (copulation solicitation), with intermediate values given for more moderate signs of interest (paying attention, beak wiping, hopping in courtship display with head, and tail bent towards the male). Depending on the intensity and duration of such signals, a score to the nearest 0.1 was given to each of the 2 females for each 10 min period of the trial (realized range of scores 0–0.8). The average scores for each female over the two 10 min periods showed an individual repeatability of 0.44 across the (usually) 12 tests per female ($n = 452$ scores, $n = 40$ females). For each 20 min test, we calculated the difference in responsiveness between the high- and low-fecundity female and assessed its explanatory value for the proportion of time males associated with the high-fecundity female.

Two-way choice tests: nest-building

To study how often males would actually end up paired to the high-fecundity female when allowed enough time to choose, we conducted another choice experiment where males were given the opportunity (a 10-day period) to build a nest for each of the 2 females. For this purpose, we added 2 nest boxes on each side of 2-way choice chamber, 1 accessible to the male only, 1 accessible to the female only (Supplementary Figures S1B, S2). Both sexes had used these boxes in the aviary breeding experiment. This setup allowed the potential partners to sit next to each other, initially separated only by wires, but at a later stage—after the male built a nest—also by nest material (see Supplementary Figure S2). The bottom of all nest boxes had been filled with hay before the start of the experiment and each male had access to coconut fibres in the neutral zone to build a nest. Every day of the 10-day experiment we recorded the approximate number of coconut fibres in each of the 2 nests of the male, as well as the number of eggs in each of the females' nests. Male preference was scored on a daily basis according to nest size (judged by the difference in total accumulated fibres in the 2 nests). However, for analysis, we scored whether the high-fecundity female was chosen (referred to as the binomial variable "correct choice": 0 or 1) based on the relative nest size on the day before the first egg was laid (by 1 of the 2 females or on day 10 if no eggs were laid).

Each female duo ($n = 8$, $n = 10$ in replicate 1 and 2) was tested with 3 out of the 6 males that participated in the previous choice chamber experiment (always choosing 1 randomly from each week). Tests were done in the same order as before (e.g., A, C, F) but with a 10–16 days break in between tests with successive males. Up to 10 duos were tested simultaneously. In 52 out 54 trials, males built nests before 1 of the females started laying and at least 1 female laid an egg in 49 out of 54 trials.

Human rating of female fecundity

Potentially, there might be obvious differences (e.g., in plumage condition or visible signs of sickness) between the selected high- and low-fecundity females. To investigate whether high-fecundity females differed from low-fecundity females in any way that is obvious to naive human observers, we asked, for each replicate experiment, 52 people from our institute to guess which female of each duo is the high-fecundity one. This was done immediately after the 2-way choice experiments (Supplementary Table S2). Because 2 females had died during the choice chamber tests, there were 8 duos to be judged in replicate 1 and 10 duos in replicate 2. The order of judging (the order in which individual observers rated the duos) was randomized for each observer and fitted as a covariate in the model (to control for the possible effect that people might become better at judging over time). For judging, 2 females of each duo were randomly housed in a cage with 2 halves separated by wire mesh. People were asked to write down whether the high-fecundity female was on the left or the right side of the mesh. Across the 2 replicates, we obtained 936 guesses (replicate 1: 52 observers × 8 duos, replicate 2: 52 × 10) from 77 different observers (some participated in both replicates).

Data analysis

Data were analyzed with mixed effect models using the lme4 package (Douglas Bates 2015) in R 3.2.3 (R Core Team 2015). Male, female or human observer identity, and female duo identity were included as random intercepts and the dependent variable was modeled as a Gaussian trait (number of eggs laid, proportion of time) or binomial trait ("correct" choices). Random effect estimates were used to calculate individual repeatabilities in the presence of fixed effects (except for some fixed effects that were excluded from the final model because trends were against the expectation). Repeatability of female fecundity was calculated from the variance component of female identity divided by the total variance (female ID + residual). Likewise, repeatability of male time allocation and effects of "female duo" on "male time allocation to the high fecundity female" were calculated in the same way: male ID/(male ID + female duo ID + residual); female duo ID/(male ID + female duo ID + residual). The significance of random effects was assessed by comparing models with and without the respective random effect using a likelihood ratio test. The resulting P values were divided by 2 (Bolker et al. 2009), because when the null hypothesis is true, this test yields "$P = 1$" in 50% of the cases where an Anova would yield $P > 0.5$. P values for fixed effects were calculated from t values (with infinite df) when the model output did not provide P values directly. Confidence intervals were calculated as estimate ± 2SE. For all results from choice tests, we report parameter estimates from models rather than averages or proportions calculated from the raw data (e.g., proportion of "correct" choices) because parameter estimates account for the nonindependence of data points. We present all results irrespective of significance and all analysis decisions were made independent of significance (unbiased reporting). Furthermore, in response to requests to assess the replicability of research findings (Freedman et al. 2015), we present our results separately for each of the 2 replicates. For the assessment of effect sizes, which are often small in evolutionary ecology (Jennions and Møller 2002), we also present a joint analysis of the replicates.

RESULTS

Fecundity of selected females

The models describing variation in female fecundity in the 2 replicate breeding experiments are shown in Supplementary Table S3. After accounting for the social pairing situation, variation in female fecundity was mostly due to differences in readiness to initiate a full clutch (low-fecundity females were less likely to start laying and if they laid eggs, they rarely produced a full clutch, see Supplementary Figure S3). In replicate 1, the number of eggs laid by individual females was moderately repeatable ($R = 0.45$, $N = 107$ females) and depended on the social pairing situation of the female. Females that were paired for longer and that were bonded more exclusively laid more eggs (see "days paired" and "female share" in Supplementary Table S3). In replicate 2, the repeatability of female fecundity was slightly lower ($R = 0.32$, $N = 106$). The number of eggs laid by females depended on the number of days a female was socially paired but not on the exclusiveness of the pair bond (Supplementary Table S3).

Across the 2 replicates, females laid a total of 4041 eggs, 95% of which had been assigned to social parents based on nest attendance (3215 fertile and 625 apparently infertile eggs). Genetic parentage analysis revealed that 222 out of 3215 eggs (6.9%) had been wrongly assigned to a female (a result of egg dumping or take-over of nests). All fertile eggs for which we previously had no assignment to a social mother ($n = 141$ eggs) were successfully assigned to their genetic mothers based on the molecular data. Taking this into account, differences in "true fecundity" between the selected top and bottom females (based on "estimated fecundity", see Methods for details) were somewhat reduced (Table 1), as already expected from regression to the mean (Barnett et al. 2005; Kelly and Price 2005). Still, high-fecundity females had laid about 5 times more eggs than low-fecundity females (replicate 1: 5.7 times, replicate 2: 4.8 times). Moreover, differences in fecundity were confirmed in the later nest-building experiment, where, despite regression toward the mean, high-fecundity females laid more eggs than low-fecundity females (replicate 1: 6.6 times, replicate 2: 2.8 times).

The high-fecundity females were on average slightly heavier (body mass: mean ± SD = 16.44 ± 1.67 g) than the low-fecundity females (15.54 ± 1.51 g; paired t-test on $n = 20$ duos: $t = 2.30$, df = 19, $P = 0.033$) but the groups did not differ in beak coloration scores (high-fecundity: −1.19 ± 0.87, low-fecundity: −1.25 ± 0.63; $t = 0.23$, df = 19, $P = 0.82$) or in average responsiveness scores during choice trials (high-fecundity: 0.48 ± 0.11, low-fecundity: 0.46 ± 0.12; $t = 0.60$, df = 19, $P = 0.55$) and also not in age when tested for fecundity (high-fecundity: 602 ± 165 days, low-fecundity: 658 ± 165 days; $t = -1.38$, df = 19, $P = 0.18$).

Two-way choice tests: male time spent near females

The proportion of time that male zebra finches associated with the high-fecundity female of a duo during the 2-way choice tests was normally distributed (Figure 1). In replicate 1, males spent on average 55.1% of their active time with the high-fecundity female, which significantly differed from random ($n = 104$ choice tests, $n = 53$ males, $n = 10$ female duos, $P = 0.002$, Supplementary Table S4). Moreover, this proportion increased significantly (by 9.0%) from the first to the second test of each male ($P = 0.0002$, Supplementary Table S4), suggesting that males were better at selecting the high-fecundity female on their second test day. None of the female characteristics (beak color, body mass,

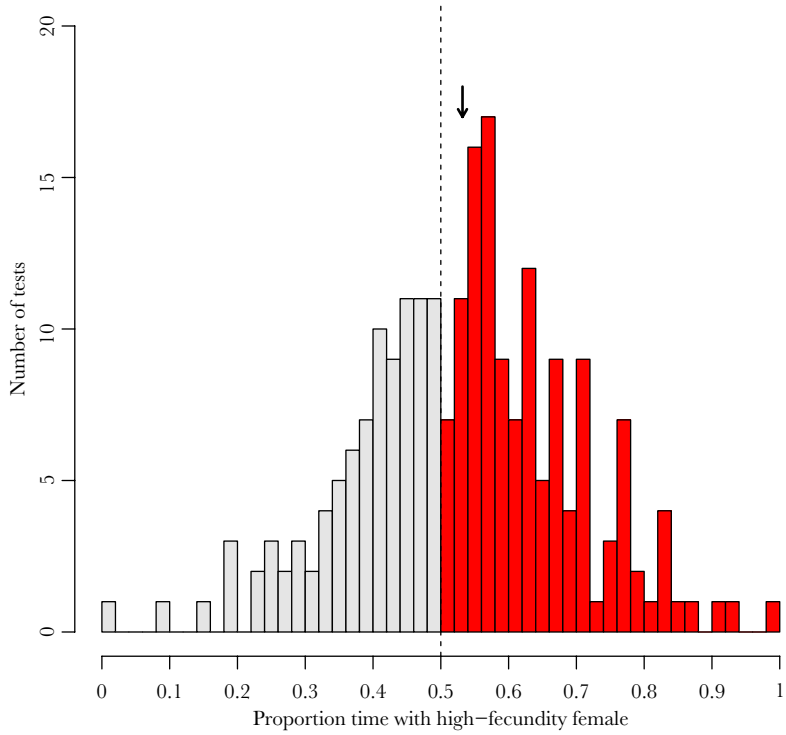


Figure 1
Histogram of the relative time that males associated with the high-fecundity female during the 2-way choice tests. Proportion of time higher than 0.5 (above dotted line in red) means that males spent more active time with the high-fecundity female than with the low-fecundity female ("correct choices"). The *y* axis shows the number of choice tests (n = 221, each lasting 20 min). The arrow indicates the estimated intercept (population average: 0.53) from a mixed effect model (see Table 2).

Table 2

Linear mixed model explaining the proportion of time that males associate with the high-fecundity female (replicate 1 and 2 combined)

	Estimate (β ± SE)	T	*P*	Repeatability
Random effects:				
Male ID (n = 113)	0.0027			0.11
Female duo ID (n = 20)	0.0013			0.05
Fixed effects:				
Intercept	0.532 ± 0.014	2.34	0.019	
Replicate (2nd vs. 1st)	−0.032 ± 0.028	−1.15	0.250	
Male test order (2nd vs. 1st)	0.045 ± 0.019	2.30	0.021	
Mass difference	0.005 ± 0.008	0.56	0.58	
Beak color difference[a]	−0.005 ± 0.012	−0.44	0.66	
Responsiveness difference[a]	−0.038 ± 0.053	−0.72	0.47	
Age difference (yrs)[a]	0.051 ± 0.026	1.96	0.05	
Male choosing motivation[a,b]	−0.011 ± 0.012	−0.99	0.32	

The intercept is tested against 50% (random choice). All fixed effects were mean centred; hence the intercept refers to the average or intermediate condition of covariates and factors.
[a]Covariates excluded from the final version of the model, because the trends were opposite to expectations (suggesting preferences for less red females, less responsive females, older females, and better choices by less motivated males). Inclusion of these covariates has only minimal effects on other parameter estimates. Not included and not shown is a further, post hoc test for male preferences for females with intermediate beak color (estimate ± SE: 0.014 ± 0.026, t = 0.56, P = 0.58; trend opposite to expectation).
[b]The total time the male spent courting or paying attention to any of the 2 females.

age, responsiveness) explained variation in male preference (Supplementary Table S4). Males that spent more time associating with any of the females ("choosing motivation") did not show stronger preferences for the high-fecundity female (Supplementary Table S4). Individual males were repeatable in their choice across the 2 tests (R = 0.27) but the 10 female duos did not differ consistently in "male time allocation to the high fecundity female" across the 6 test males (R = 0, Supplementary Table S4).

In replicate 2, the average time spent with the high-fecundity female was 51.4%, which did not differ significantly from 50% (n = 117 choice trials, n = 60 males, n = 10 duos, P = 0.52, Supplementary Table S4). Also, this proportion did not increase from the first to the second test (P = 0.86, Supplementary Table S4). Again, male preferences were neither explained by female characteristics nor by male choosing motivation (Supplementary Table S4), with the exception of a significant male preference for older females. However, we excluded female age as a predictor from the final model, because the estimate was opposite to expectations (younger females show higher fecundity; see Methods for details). Inclusion of this factor did not alter any of the conclusions drawn from the model. In the second replicate, male choice of females was not repeatable (R = 0) and the 10 female duos differed only slightly in "male time allocation to the high fecundity female" (R = 0.09).

A joint analysis of the 2 replicates yielded a weak but significant male preference for high-fecundity females (P = 0.019, Table 2). In this model, we added "replicate" (2 levels) as another fixed effect. The model revealed that some of the variance was explained by

the random effect "female duo ID" (Table 2), suggesting that male choice for the high-fecundity female was somewhat stronger in some duos than in others (yet this variance component was not significant judging from the change in Akaike Information Criterion, $\Delta AIC = 1.45$, $P = 0.23$).

We also analyzed male choice as a binary trait ("correct" if relative time with high-fecundity female is > 50%, in a mixed effect model controlling for the same random effects as in Table 2). In replicate 1, males chose the "correct" in 64% of the cases ($P = 0.019$, 95% confidence interval [CI] 52–74%) and in replicate 2, male choice was "correct" female in 56% of the cases ($P = 0.32$, 95% CI 44–67%). Across both replicates, the proportion of "correct" choices was 59% ($P = 0.013$, CI 52–66%).

Two-way choice tests: male nest-building

On average, "final choice decisions" (based on the difference in coconut fibres in the 2 nests) were recorded after 4.5 days (SD = 2.2, range 1–9 days). In replicate 1, males built the biggest nest (see Supplementary Figure S2) for the high-fecundity female in 62% of the cases (n = 23 males, n = 8 duos, $P = 0.33$, 95% CI 38–81%). In replicate 2, the proportion of "correct" choice was 60% (n = 29 males, n = 10 duos, $P = 0.42$, 95% CI 35–81%). Overall, males preferentially built a nest for the high-fecundity female in 61% of cases (n = 52, $P = 0.21$, 95% CI 44–76%). In the respective mixed-effect model with binomial error structure, the random effect of female duo was not significant (model on both replicates: $\Delta AIC = 1.42$, $P = 0.22$). In this experiment, high-fecundity females laid more eggs than low-fecundity females (see Table 1), indicating that males would indeed have benefited from pairing with the high-fecundity female rather than the low-fecundity one.

Human rating of fecundity

We analyzed human responses (binary variable, 1 = correct guess of the high-fecundity female, 0 = incorrect guess, chance probability = 50%) in a mixed effect model with observer identity and female duo identity as random effects. Human guesses were correct in 59.3% of the cases in replicate 1 (n = 416 judgments, n = 52 observers, n = 8 duos, $P = 0.088$, 95% CI 48–69%) and in 57.3% in replicate 2 (n = 520 judgments, n = 52 observers, n = 10 duos, $P = 0.21$, 95% CI 46–68%). Across the 2 replicates, the rate of correct judgment was 58.2% (n = 936 judgments, n = 77 observers, n = 18 duos, $P = 0.043$, 95% CI 50–66%). In this model, the random effect "observer" did not explain any variance (i.e., observers did not differ in their abilities to identify the high-fecundity female) but the random effect of "female duo" had a large effect on human ratings ($\Delta AIC = 48.1$, $P < 10^{-11}$).

We also compared the human rating and the choices made by zebra finch males of these 18 female duos (Pearson $r = -0.16$, $N = 18$ duos, $P = 0.53$, Supplementary Figure S4). This means that there was no "consensus" between human ratings and the choices made by zebra finch males.

DISCUSSION

Limited male abilities

Figure 2 summarizes all tests (2-way choice tests: male time spent near females; male nest-building and human rating) in the form of "proportion correct choices." Across the 2 replicates, only 1 out of 8 tests reached statistical significance and the proportion of "correct" choices does not seem to increase notably when males are given more time to make their choice (from day 1 to day 2 and to nest building).

Our 2-way choice experiments reveal a slight but significant tendency for males to preferentially associate with the higher-fecundity female (Figure 1, Table 2). When males are given enough time to select a partner (i.e., built a nest for 1 of the 2 females), this bias in favor of high-fecundity females appears to be the strongest (61% "correct choices") but due to a more limited sample size, the 95% confidence interval around this estimate remained rather wide (44–76%). Human observers also demonstrated a significant ability to identify the high-fecundity individual correctly but the rate of correct guesses was not high (58%).

Depending on perspective, the question whether male zebra finches can assess differences in female fecundity can now be answered with either "yes" or "no": "yes" in the sense that we found a statistically detectable effect when averaging among a large number of tests. However, the answer is "no" in the sense that many males still spent more time or built a nest with the low-fecundity female, even though the experimental design maximized the contrast between the 2 females, by picking individuals of high versus low fecundity. In principle, males could have picked up either cues that distinguish females of top fecundity from the population average (high-fecundity females laid 5–7 eggs more than the population mean; see BLUPs in Table 1) or cues that identify females in really poor condition (low-fecundity females laid 6–9 eggs less than the population mean; Table 1) but apparently neither of these hypothetical cues seem to allow males to reliably choose the better option. Using the BLUPs from our models as a guideline, any randomly chosen duo of females will differ on average by 4.9 eggs, whereas the females we selected differed on average by 13.5 eggs. Thus, at population level, when differences in fecundity become less extreme, we expect male choice to become even less accurate. It is not clear whether variance in female fecundity in our domesticated population is relatively high or low compared to the wild. In our population, variance in inbreeding contributes substantially to variance in fecundity (Forstmeier et al. 2012), while inbreeding is almost completely absent in the wild (Knief et al. 2015). In the wild, environmental stressors may play an additional role but it seems unlikely that males in the wild would often choose between females that differ about four-fold in fecundity (as in our experiment; Table 1 last column). Hence, if the ability to detect differences declines with the magnitude of the difference, we expect that the realized average benefits of male choosiness will be fairly small overall.

Comparison of replicates

Given that many published results do not seem to be robust (Ioannidis 2005; Open-Science-Collaboration 2015), replicability of research findings is currently a hot topic of debate (Freedman et al. 2015). In evolutionary ecology, low replicability may be expected because realized effect sizes tend to be small (Jennions and Møller 2002), so that we typically lack the power to detect these effects (Parker et al. 2016). The current study can be seen as an example of this situation. The average effect that we describe is modest and rarely reaches significance in a single test (Figures 1 and 2). On the one hand, one could say that replicate 2 represents a failure to confirm the findings made in replicate 1, because both the significant intercept and the significant order effect largely disappeared (Supplementary Table S4). On the other hand, all

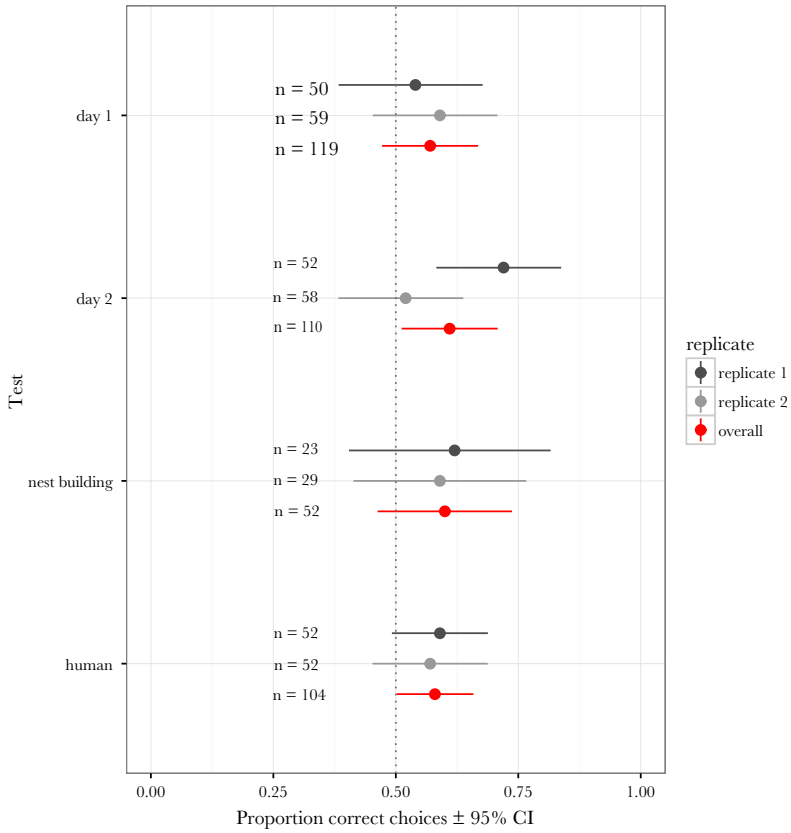


Figure 2
Summary of all tests in the form of "proportion correct choices ± 95% CI" across the 2 replicate experiments. "day 1" and "day 2" refer to the time males spent near the females in choice chamber tests on 2 different days involving the same set of males. "Nest building" refers to the nest-building behavior in the 2-way choice test (involving a subset of the same males) where male choice was recorded before egg-laying. The last test shows the proportion of "correct choices" by human observers. Sample sizes refer to numbers of males with informative trials or to the number of human observers.

observed trends in replicate 2 were in the expected direction, and the estimates are not substantially different from those of replicate 1 (see Figure 2). Moreover, the weaker effect of male choice for high-fecundity females in replicate 2 (compared to replicate 1) corresponds with less favorable conditions for detecting the effect. In replicate 2, the repeatability of female fecundity was lower than in replicate 1 (R = 0.32 vs. R = 0.45) and the difference between high- and low-fecundity females in how many eggs they laid during the nest-building experiment was also less pronounced in replicate 2 compared to replicate 1 (see last column in Table 1). This reiterates the point that male choice should become less accurate as the between-female difference in fecundity becomes smaller.

Absence of reliable cues

Assuming that male preference for high-fecundity females exists, the cues that males have used for the identification of these females remain obscure. High-fecundity females did not signal a higher "readiness to breed" by the use of positive courtship signals like ritualized body postures (bending of the tail, copulation solicitation). The selected high- and low-fecundity females also did not differ in beak color and in the entire data set (2 rounds of aviary breeding) there was no correlation between the total number of eggs laid and beak color (scored on the Munsell scale [Bolund et al. 2007] at the age of reaching maturity day 100–120; r = 0.01, P = 0.84, n = 213 females). The total number of eggs laid during the fecundity experiment was related to female age (older females were less fecund) but males did not prefer younger females in choice tests. Most notably, the selected high-fecundity females were significantly heavier than the low-fecundity females and also in the entire data set there was a significant positive correlation between female fecundity and body mass at 100–120 days of age (n = 213, r = 0.20, P = 0.003). Hence, males could use body mass as a cue to female fecundity but it would be a cue that only explains about 4% of the variation in fecundity. This is in line with other studies on birds, where female body mass is only a weak predictor of clutch size (Potti 1999; Haywood 2013;). In our zebra finches, female mass at reaching maturity showed a coefficient of variation (CV) of only 9.3% (n = 213, body mass mean ± SD = 15.41 ± 1.43 g), which is similar to zebra finches in the wild (body mass mean ± SD = 12.44 ± 0.98 g, CV = 7.9%); data from Knief et al. 2016, whereas coefficients of variation can be much higher in reptiles (e.g., 21%, Bjorndal et al. 2013) or insects (e.g., 24.5%, Calvo and Molina 2005) where males show clear preferences for heavier females (e.g., Edward and Chapman 2012) and heavier females indeed lay more eggs (see Introduction for details). In our choice tests, males did not seem to respond to either female responsiveness, female beak color, potential indicators of female age, or female body mass (Table 2, Supplementary Table S4). Also, Rutstein et al. 2007 did not find that male zebra finches preferentially courted larger females when mixing domesticated and wild birds, which differ markedly in body size (Forstmeier et al. 2007).

Would males benefit from preferring heavy females?

If males had reliable cues to identify the high-fecundity female, they likely would have profited from choosing her. As the nest-building experiment showed, the selected high-fecundity females indeed laid many more eggs than the low-fecundity females (Table 1), irrespective of earlier investment in egg-laying. Hence, we assume that males were largely unaware of the intrinsic differences in fecundity of the presented females. The most reliable indicator of female fecundity that we identified was female body mass (see above). However, if female mass explains only about 4% of the variance in female fecundity, it probably explains even less of the variance in male fitness in the wild, given that hatching failure and nest predation add noise to the relationship between the body mass of the male's social partner and his lifetime reproductive success. It therefore seems implausible that the small benefits of preferring heavy females would offset the costs of increased male–male competition for the heavier females. This might explain why male zebra finches apparently did not evolve such directional preferences for heavy females that would lead to consensus among males regarding female attractiveness.

The observed outcome of male choice tests (Table 2, Supplementary Table S4) is somewhat similar to our experience with female choice tests (Forstmeier and Birkhead 2004; Schielzeth et al. 2010). Females show only a modest repeatability in their individual preferences when tested with the same set of males and females show remarkably little consensus in who they prefer (i.e., low repeatability of male attractiveness; Forstmeier and Birkhead 2004). In this study, the test males also showed a low repeatability in their individual preferences when tested 2 days apart ($R = 0.11$, Table 2) and there was little between-male agreement on whether the high-fecundity female of a duo was attractive or not ($R = 0.05$, Table 2). Hence, measuring population-wide preferences in choice tests remains a challenge because the extent of consensus among individuals is very limited (Forstmeier and Birkhead 2004).

In conclusion, our experiments revealed a significant but limited ability of males to select the more fecund female when given a choice between 2 extremes. When given sufficient time for choosing a partner, male success in pairing with the high-fecundity females was only 61% (95% CI 44–76%). Given that the upper limit of the confidence interval lies at 76% of "correct" choices, we can confidently say that male abilities to choose highly fecund females are far from perfect even when confronted with females that differ substantially in fecundity.

SUPPLEMENTARY MATERIAL

Supplementary data are available at *Behavioral Ecology* online.

FUNDING

This work was supported by the Max Planck Society (to B.K.) and the China Scholarship Council (CSC; stipend to D.W.).

The authors thank Katrin Martin for extensive help with the breeding experiments and for help with setting up choice trials, Mercedes Diesel for help with running choice trials, and Melanie Schneider for molecular work. The authors also thank Sonja Bauer, Andrea Kortner, Jane Didsbury, and Petra Neubauer for the animal care.

Data accessibility: Analyses reported in this article can be reproduced using the data provided by Wang et al. (2017).

Handling editor: Naomi Langmore

REFERENCES

Alonzo SH. 2012. Sexual selection favours male parental care, when females can choose. Proc Biol Sci. 279:1784–1790.
Amundsen T. 2000. Why are female birds ornamented? Trends Ecol Evol. 15:149–155.
Andersson M, Simmons LW. 2006. Sexual selection and mate choice. Trends Ecol Evol. 21:296–302.
Andersson MB. 1994. Sexual selection. Princeton (NJ): Princeton University Press.
Arntzen JW. 1999. Sexual selection and male mate choice in the common toad, Bufo bufo. Ethol Ecol Evol. 11:407–414.
Barnett AG, van der Pols JC, Dobson AJ. 2005. Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 34:215–220.
Bjorndal KA, Parsons J, Mustin W, Bolten AB. 2013. Threshold to maturity in a long-lived reptile: interactions of age, size, and growth. Mar Biol. 160:607–616.
Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MH, White JS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 24:127–135.
Bolund E, Martin K, Kempenaers B, Forstmeier W. 2010. Inbreeding depression of sexually selected traits and attractiveness in the zebra finch. Anim Behav. 79:947–955.
Bolund E, Schielzeth H, Forstmeier W. 2007. Intrasexual competition in zebra finches, the role of beak colour and body size. Anim Behav. 74:715–724.
Bolund E, Schielzeth H, Forstmeier W. 2010. No heightened condition dependence of zebra finch ornaments–a quantitative genetic approach. J Evol Biol. 23:586–597.
Bonduriansky R. 2001. The evolution of male mate choice in insects: a synthesis of ideas and evidence. Biol Rev Camb Philos Soc. 76:305–339.
Calvo D, Molina JM. 2005. Fecundity-body size relationship and other reproductive aspects of Streblote panda (Lepidoptera: Lasiocampidae). Ann Entomol Soc Am. 98:191–196.
Charmantier A, Sheldon BC. 2006. Testing genetic models of mate choice evolution in the wild. Trends Ecol Evol. 21:417–419.
Clutton-Brock TH, Parker GA. 1992. Potential reproductive rates and the operation of sexual selection. Q Rev Biol. 67:437–456.
Cote IM, Hunte W. 1989. Male and female mate choice in the redlip blenny - why bigger is better. Anim Behav. 38:78–88.
Douglas Bates MM, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. J Stat Softw. 67:1–48.
Edward DA, Chapman T. 2011. The evolution and significance of male mate choice. Trends Ecol Evol. 26:647–654.
Edward DA, Chapman T. 2012. Measuring the fitness benefits of male mate choice in Drosophila melanogaster. Evolution. 66:2646–2653.
Forstmeier W, Birkhead TR. 2004. Repeatability of mate choice in the zebra finch: consistency within and between females. Anim Behav. 68:1017–1028.
Forstmeier W, Schielzeth H, Mueller JC, Ellegren H, Kempenaers B. 2012. Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. Mol Ecol. 21:3237–3249.
Forstmeier W, Segelbacher G, Mueller JC, Kempenaers B. 2007. Genetic variation and differentiation in captive and wild zebra finches (*Taeniopygia guttata*). Mol Ecol. 16:4039–4050.
Freedman LP, Cockburn IM, Simcoe TS. 2015. The economics of reproducibility in preclinical research. PLoS Biol. 13:e1002165.
Griggio M, Devigili A, Hoi H, Pilastro A. 2009. Female ornamentation and directional male mate preference in the rock sparrow. Behav Ecol. 20:1072–1078.
Haywood S. 2013. Origin of evolutionary change in avian clutch size. Biol Rev Camb Philos Soc. 88:895–911.
Hoelzer GA. 1989. The good parent process of sexual selection. Anim Behav. 38:1067–1078.
Houde AE, Endler JA. 1990. Correlated evolution of female mating preferences and male color patterns in the guppy *Poecilia reticulata*. Science. 248:1405–1408.
Ihle M, Forstmeier W. 2013. Revisiting the evidence for inbreeding avoidance in zebra finches. Behav Ecol. 24:1356–1362.
Ioannidis JPA. 2005. Why most published research findings are false. Plos Med. 2:696–701.
Jennions MD, Møller AP. 2002. Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. Biol Rev Camb Philos Soc. 77:211–222.

Jensen H, Saether BE, Ringsby TH, Tufto J, Griffith SC, Ellegren H. 2004. Lifetime reproductive success in relation to morphology in the house sparrow Passer domesticus. J Anim Ecol. 73:599–611.
Jones IL, Hunter FM. 1993. Mutual sexual selection in a monogamous seabird. Nature 362:238–239.
Jones KM, Monaghan P, Nager RG. 2001. Male mate choice and female fecundity in zebra finches. Anim Behav. 62:1021–1026.
Kelly C, Price TD. 2005. Correcting for regression to the mean in behavior and ecology. Am Nat. 166:700–707.
Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W. 2015. Quantifying realized inbreeding in wild and captive animal populations. Heredity (Edinb). 114:397–403.
Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W. 2016. Fitness consequences of polymorphic inversions in the zebra finch genome. Genome Biol. 17:199.
Koops MA, Hutchings JA, McIntyre TM. 2004. Testing hypotheses about fecundity, body size and maternal condition in fishes. Fish and Fisheries 5:120–130.
Long TA, Pischedda A, Stewart AD, Rice WR. 2009. A cost of sexual attractiveness to high-fitness females. PLoS Biol. 7:e1000254.
Mautz BS, Møller AP, Jennions MD. 2013. Do male secondary sexual characters signal ejaculate quality? A meta-analysis. Biol Rev Camb Philos Soc. 88:669–682.
Monaghan P, Metcalfe NB, Houston DC. 1996. Male finches selectively pair with fecund females. P Roy Soc B-Biol Sci. 263:1183–1186.
Nakagawa S, Parker TH. 2015. Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. BMC Biol. 13:88.
Open-Science-Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349:aac4716-1-8. doi: ARTN aac4716.10.1126/science.aac4716.
Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S. 2016. Transparency in ecology and evolution: real problems, real solutions. Trends Ecol Evol. 31:711–719.
Pelabon C, Borg AA, Bjelvenmark J, Forsgren E, Barber I, Amundsen T. 2003. Do male two-spotted gobies prefer large fecund females? Behav Ecol. 14:787–792.
Potti J. 1999. Maternal effects and the pervasive impact of nestling history on egg size in a passerine bird. Evolution 53:279–285.
R Core Team. 2015. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. http://www.R-project.org/.
Rutstein AN, Brazill-Boast J, Griffith SC. 2007. Evaluating mate choice in the zebra finch. Anim Behav 74:1277–1284.
Schielzeth H, Bolund E, Forstmeier W. 2010. Heritability of and early environment effects on variation in mating preferences. Evolution. 64:998–1006.
Schielzeth H, Kempenaers B, Ellegren H, Forstmeier W. 2012. QTL linkage mapping of zebra finch beak color shows an oligogenic control of a sexually selected trait. Evolution. 66:18–30.
Sheldon BC. 1994. Male phenotype, fertility, and the pursuit of extra-pair copulations by female birds. P Roy Soc B-Biol Sci 257:25–30.
Smiseth PT, Amundsen T. 2000. Does female plumage coloration signal parental quality? A male removal experiment with the bluethroat (Luscinia s. svecica). Behav Ecol Sociobiol. 47:205–212.
Swierk L, Myers A, Langkilde T. 2013. Male mate preference is influenced by both female behaviour and morphology. Anim Behav. 85:1451–1457.
Torres R, Velando A. 2005. Male preference for female foot colour in the socially monogamous blue-footed booby, *Sula nebouxii*. Anim Behav. 69:59–65.
Trivers RL. 1972. Parental investment and sexual selection. In: Campbell B, editor. Sexual selection and the descent of man. Chicago (IL): Aldine.
von Schantz T, Bensch S, Grahn M, Hasselquist D, Wittzell H. 1999. Good genes, oxidative stress and condition-dependent sexual signals. Proc Biol Sci. 266:1–12.
Wang D, Kempenaers N, Kempenaers B, Forstmeier W. 2017. Data from: male zebra finches have limited ability to identify high-fecundity females. Dryad Digital Repository. http://doi:10.5061/dryad.vc0hv.
Willemsen RE, Hailey A. 1999. Variation of adult body size of the tortoise Testudo hermanni in Greece: proximate and ultimate causes. J Zool. 248:379–396.
Zahavi A. 1977. The cost of honesty (further remarks on the handicap principle). J Theor Biol. 67:603–605.
Zedrosser A, Dahle B, Swenson JE. 2006. Population density and food conditions determine adult female body size in brown bears. J Mammal. 87:510–518.

Male zebra finches have limited ability to identify high-fecundity females

Daiping Wang, Nele Kempenaers, Bart Kempenaers, Wolfgang Forstmeier

SUPPLEMENT

Supplementary Methods: Study population and assessment of fecundity

Subjects of the current study are from a population of captive zebra finches maintained at the Max Planck Institute for Ornithology in Seewiesen, Germany (Forstmeier et al., 2007) (population #18 in Forstmeier et al. 2007). Housing conditions, diet and aviary specifications for breeding were described in detail in (Schielzeth et al., 2010). For the last three generations, this population has been split into six selection lines that were bred for high versus low courtship rate (2 high lines, 2 unselected control lines, 2 low lines; see (Mathot et al., 2013)). We here focus on the entire third generation of these selection lines (irrespective of line because we have no indications that lines differ in either female fecundity or female attractiveness), initially consisting of 681 birds hatched between July 2012 and June 2013.

For the purpose of another study we assessed the frequency of extra-pair paternity for these birds between January 2014 and May 2015, which yielded estimates of female fecundity as a by-product. Breeding was organized as follows: birds were randomly (irrespective of age) assigned to four successive groups each comprising 54 males and 54 females (216 of each sex in

total). Each group was then assigned to one of nine aviaries, such that (1) each aviary contained one male and one female from each selection line (9 aviaries x 6 lines corresponding to 54 individuals of each sex) and (2) all birds within an aviary were unfamiliar with each other. The colour-banded birds (colours: white, yellow, orange, light blue, blue, black) could freely choose a mate and laid to up to three clutches within a period of seven weeks (nest boxes were provided from day 1 to day 45). All eggs laid were replaced by plastic eggs as soon as found and collected for later parentage assignment. Clutches consisting of plastic eggs were removed after 10 days of incubation to allow the female to lay the next clutch. On day 49, individuals were separated by sex into different rooms for a two-week period, after which we initiated an identical second round of breeding, but with a different set of potential partners (by swapping the six males of one aviary to the next). This allowed us to quantify the repeatability of female fecundity, estimated as total number of eggs laid, with different male partners. During the breeding experiment, we strictly provide the birds with standard food. That is: birds generally always have access to ad libitum drinking water, cuttlefish bone for calcium supply and grit to aid digestion. Moreover, they receive salad leaves and a multivitamin solution once a week. The main diet of our zebra finches consists of a mixture of six kinds of seeds (29% pearl white millet, 29% panicum millet, 14% Japanese millet, 14% canary seed, 7% Dakota red millet, 7% yellow millet).

During the 2 x 7 weeks of breeding (Table S2), we observed all birds daily for approximately 30 min each time (about 120 times in total) and recorded all instances of allopreening, sitting in body contact, and visiting a nest-box together. From these observations we extracted two parameters of pairing success (1) the number of days (out of 49) that a female was socially

paired ('days paired': mean for paired females ± SD = 43.7 ± 9.5 days, median = 48 days, n = 372; unpaired females n = 54 were given a score of zero). For this parameter, we defined the start of pairing as the first evidence of 'exclusive' bonding by the female to one male (i.e. >50% of bonding behaviours directed to one male; minimum 8 observations on this female-male combination). (2) The exclusivity of her partner showing such pair bonding behaviour only with her, calculated as the proportion of records of the above behaviours by her partner that was directed to her ('female share': mean ± SD = 0.88 ± 0.17, median = 0.96, n = 372; unpaired females n = 54 were given a score of zero). Daily nest checks combined with behavioural observations allowed us to assign 95% of all eggs laid (3840 out of 4041) to social parents that attended the respective nest. Based on this social assignment, we recorded the putative number of eggs laid by each female in each round ('estimated fecundity'). Because 201 eggs were not assigned and because females may also lay eggs into a nest owned by another female ('egg dumping'; 5.4% of all eggs in an earlier study, (Schielzeth et al., 2010)), we used molecular parentage analysis of 3356 fertile eggs plus social assignment of 625 infertile eggs (which were not genotyped) to assess 'true fecundity' (for the remaining 60 infertile eggs there was no social assignment of parents). However, this information became available only after the choice tests (see below). For parentage analysis, we used 15 highly polymorphic microsatellite markers (Table S1), which allows a practically error-free assignment of all offspring (given that extensive SNP genotyping (Backstrom et al., 2010) had confirmed error-free assignment in previous work).

Because the time required for the measurement of fecundity created a lag between successive groups, we carried out two practically identical replicate experiments with extreme females chosen from a pool of 2 x 54 (= 108) females. The first two groups were breeding from January

to September 2014, comprising 107 females that participated in both rounds (one female died in round 1), and these were used for replicate 1 of our experiments. The groups three and four were breeding from October 2014 to May 2015, comprising 106 females that participated in both rounds (two females died in round 1), and these were used for replicate 2 of our experiments (see Table S2).

Table S1. Primers for 15 microsatellite markers and PCR conditions used for parentage assignment. DNA was extracted from tissue samples using DNeasy Blood & Tissue Kit (Qiagen).The Type-it Microsatellite PCR Kit (Qiagen) was used for genotyping. Each 10µl PCR reaction contained 20-200 ng DNA, 5µl of 2x Qiagen Type-it master mix, 3µl of H20 and 1µl of primer mix. Each 1.5µl PCR product was analysed on a ABI 3130 Genetic Analyser with POP4 as polymer and GS-500 (LIZ) as size standard (all Applied Biosystems) under standard conditions.

chromosome	primer name	fluorescence label	sequence	mix	mix volume	volume of primer (stock concentration 100µM)	annealing temperature	N of cycles	remarks
Tgu1A	chr1A_39MB_F	NED	GGCTCCTTAAAAGCCCAGCTC	4	300	0.7	60°C	23	
Tgu1A	chr1A_39MB_R		CTCTGCTGGACCCTCTCTAG	4	300	0.7	60°C	23	
Tgu2	Tgu8_F	6FAM	GGGAGAGATAAAAGGTATTTTCAGG	2	400	2	57°C	21	[1]Forstmeier et al. 2007
Tgu2	Tgu8_R		GAAAGGCATGGCAATAGTGAAG	2	400	2	57°C	21	
Tgu2	Tgu2_SD44_F	VIC	TGGAAGTGGCAAGGACAACA	2	400	2	57°C	21	[2]Knief et al. 2015
Tgu2	Tgu2_SD44_R		TCCCTGCTCCCTATCTGTAT	2	400	2	57°C	21	
Tgu2	Tgu2_SD60_F	PET	CGTCCCAAAACACCAATCGT	2	400	2	57°C	21	[2]Knief et al. 2015
Tgu2	Tgu2_SD60_R		CCTCACAACACGAAGCAGAT	2	400	2	57°C	21	
Tgu3	chr3_58MB_F	PET	CCTGATTCACCATGCCCAGT	4	300	1.3	60°C	23	
Tgu3	chr3_58MB_R		AAAGGGCAGAAGGTAGACCATGA	4	300	1.3	60°C	23	
Tgu4A	chr4A_9MB_F	PET	GCCATGAACCTCTGCTCCTG	6	200	1.5	60°C	25	
Tgu4A	chr4A_9MB_R		CCACCTGCAGTGGGATTGTC	6	200	1.5	60°C	25	
Tgu5	chr5_34MB_F	PET	GCAACTGCTGCTCTGAAGGA	7	200	0.8	59°C	30	
Tgu5	chr5_34MB_R		AGCTGCACATGGGGAAGCTA	7	200	0.8	59°C	30	
Tgu6	chr6_16MB_F	VIC	TCTGCCGTGTGTGTTTCTGG	7	200	6	59°C	30	
Tgu6	chr6_16MB_R		TAGCCATCTGGGCTCCTCAA	7	200	6	59°C	30	
Tgu11	chr11_8MB_F	NED	TTGCAGGCAGGTTCAGTGTG	7	200	0.5	59°C	30	
Tgu11	chr11_8MB_R		TGGTTGCCTGGAGAAGATGG	7	200	0.5	59°C	30	
Tgu12	chr12_9MB_F	VIC	CTGTCTCACCCAGGCGAACA	6	200	0.4	60°C	25	
Tgu12	chr12_9MB_R		GCTGACTGCTCGGTTTGACC	6	200	0.4	60°C	25	
Tgu14	chr14_9MB_F	NED	GATGGAAAGGCTCTGGCACC	6	200	0.5	60°C	25	
Tgu14	chr14_9MB_R		CTGAGTGGGTCGCAGGTGAT	6	200	0.5	60°C	25	
Tgu15	chr15_6MB_F	6FAM	AGCCGAGGGCCTAAAGATGA	4	300	1.5	60°C	23	
Tgu15	chr15_6MB_R		GAGCCAGGATGAAAGGAGGT	4	300	1.5	60°C	23	
Tgu22	chr22_3MB_F	VIC	TGGCCTTGCTGACTTCTGCT	4	300	0.7	60°C	23	
Tgu22	chr22_3MB_R		AGCAGGTTGTGAGGGCTTGT	4	300	0.7	60°C	23	
Tgu26	chr26_3MB_F	6FAM	GAAAGGACCTCTGGGCTCTG	6	200	1	60°C	25	
Tgu26	chr26_3MB_R		AGCTTGCACCGTGAGGTAGC	6	200	1	60°C	25	
Tgu27	chr27_1MB_F	6FAM	GATCTGGAAATACCCTGGAGC	7	200	0.5	59°C	30	

Tgu27	chr27_1MB_R	TGAAGCATTTCCCTCTGGAGTC	7	200	0.5	59°C	30

[1] Forstmeier, W., Schielzeth, H., Schneider, M. & Kempenaers, B. 2007 Development of polymorphic microsatellite markers for the zebra finch (Taeniopygia guttata). Mol Ecol Notes 7, 1026-1028. (doi:10.1111/j.1471-8286.2007.01762.).

[2] Knief, U., Schielzeth, H., Ellegren, H., Kempenaers, B. & Forstmeier, W. 2015 A prezygotic transmission distorter acting equally in female and male zebra finches Taeniopygia guttata. Mol Ecol 24, 3846-3859. (doi:10.1111/mec.13281)

Table S2. The time schedule of experiments and respective sample sizes. During the two-way choice tests of replicate 1, two females died such that only 54 males (instead of 60) were tested.

time	experimental phase	sample sizes
replicate 1		
Jan – Sep 2014	measuring female fecundity	N = 107 females
Oct – Nov 2014	two-way choice tests	N = 10 female duos and N = 54 males
Nov 2014	human ratings	N = 8 female duos and N = 52 people
Nov 2014 – Jan 2015	two-way nest-building	N = 8 female duos and N = 24 males
replicate 2		
Oct 2014 – May 2015	measuring female fecundity	N = 106 females
Jul 2015	two-way choice tests	N = 10 female duos and N = 60 males
Aug 2015	human ratings	N = 10 female duos and N = 52 people
Aug – Oct 2015	two-way nest-building	N = 10 female duos and N = 30 males

Table S3. Linear mixed models describing variation in female fecundity (i.e. the number of eggs laid per 7-week breeding round) as a function of female identity (ID), breeding round (first vs. second), the number of days females were socially paired, and the exclusiveness of the pair bond (female share). One model is shown for each of two replicate experiments based on N = 107 and N = 106 individual females, respectively. The residuals of the models from both replicates were normal distributed (replicate 1: p = 0.10, One-Sample Kolmogorov-Smirnov Test; replicate 2: p = 0.64, One-Sample Kolmogorov-Smirnov Test). Repeatability was calculated from the variance component for female ID relative to total variance (female ID plus residual). A 'female share' value of 1 reflects exclusive pairing of the partner, while a value of 0.5 reflects equal sharing of the same male by two females. Parameter estimates refer to the number of eggs that were assigned socially to a female (before parentage analysis). From these models, BLUPs for female ID were used to select females of highest and lowest fecundity.

	estimate (β±SE)	t	*p*	repeatability
replicate1				
random effects:				
female ID	10.42			0.45
fixed effects:				
intercept	8.95±0.40	22.58	<0.001	
breeding round	-0.18±0.49	-0.36	0.720	
days paired	0.12±0.03	4.33	<0.001	
female share	4.18±1.50	2.77	0.006	
replicate2				
random effects:				
female ID	5.48			0.32
fixed effects:				
intercept	8.50±0.33	26.04	<0.001	
breeding round	0.29±0.47	0.60	0.546	
days paired	0.16±0.03	5.45	<0.001	
female share	0.81±1.45	0.55	0.582	

Table S4. Linear mixed models explaining the proportion of time that males associate with the high-fecundity female in replicate 1 and 2 separately. The intercept is tested against 50% (random choice). All fixed effects were mean centred; hence the intercept refers to the average or intermediate condition of covariates and factors.

	estimate (β±SE)	t	*p*	repeat-ability
replicate 1				
random effects:				
male ID (n = 53)	0.0058			0.27
female duo ID (n = 10)	0.0000			0.00
fixed effects:				
intercept	0.551±0.016	3.15	0.0016	
male test order (2nd versus 1st)	0.091±0.025	3.69	0.0002	
beak colour difference[a]	-0.007±0.012	-0.57	0.57	
mass difference[a]	-0.013±0.010	-1.25	0.21	
responsiveness difference	0.023±0.065	0.35	0.73	
age difference (yrs)[a]	0.003±0.038	0.09	0.93	
male choosing motivation	0.014±0.016	0.87	0.38	
replicate 2				
random effects:				
male ID (n = 60)	0.0000			0.00
female duo ID (n = 10)	0.0024			0.09
fixed effects:				
Intercept	0.514±0.021	0.65	0.52	
male test order (2nd versus 1st)	0.005±0.029	0.17	0.86	
beak colour difference[a]	-0.002±0.029	0.06	0.96	
mass difference	0.017±0.012	1.43	0.15	
responsiveness difference[a]	-0.132±0.080	-1.65	0.10	
age difference (yrs)[a]	0.080±0.036	2.24	0.03	
male choosing motivation[a]	-0.026±0.016	-1.70	0.09	

[a] Covariates excluded from the final version of the model, because the trends were opposite to expectations.

Figure S1. Schematic of the choice chamber set-up seen from above (a row of four cages, each measuring 60×40 cm and 45 cm high, two in the middle for the test male, the remaining two cages holding the females at either ends and separated from the male by wire mesh) used for the two-way choice tests (A) and for the two-way nest-building experiment (B). The male in the middle can move freely between 4 perches (black lines). During choice tests (A), we recorded the female-directed behaviour of males only when males were close to females (outside the neutral area that is indicated). For the two-way nest building experiment (B), we added four nest boxes (blue boxes), two of which were accessible for the male to build a nest for one of the females, the other two boxes allowed females to sit closely to the male and to lay eggs.

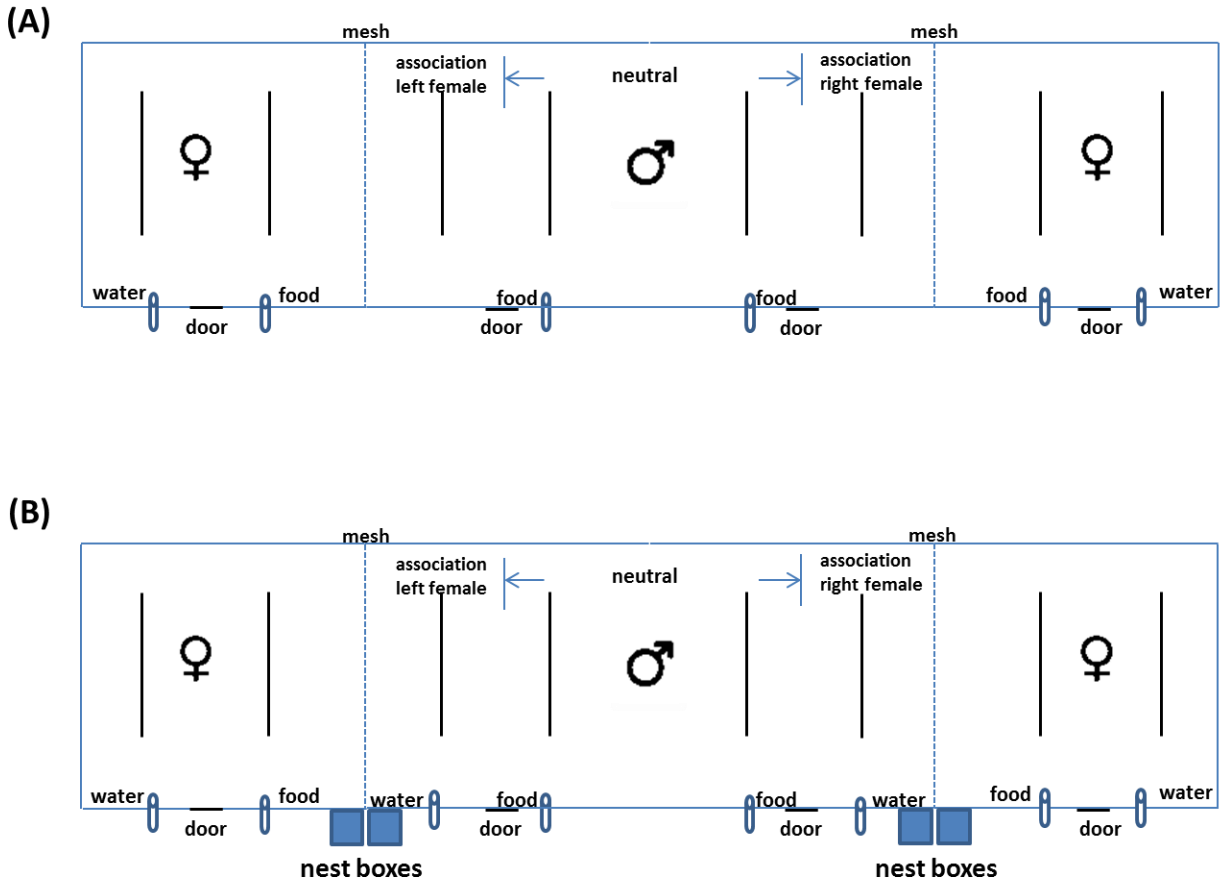


Figure S2. Photo of the two-way choice setup where males could build a nest with each female. All nest boxes were filled with some hay before the experiment, and males were given coconut fibres for nest building. Note that the male in the top row has built a nest for the female on the left (blue arrow), while the male in the bottom row has built for the female on the right (red arrow). The same apparatus was used for two-way choice tests except that no nest boxes were attached.

Figure S3. Patterns of egg laying (blue and red diamonds show eggs based on genetic assignment including socially assigned infertile eggs) and the timing of social pairing (green triangles) over the course of the two rounds of aviary breeding in high and low-fecundity females. High and low fecundity females are arranged in duos as used in the choice experiments, with corresponding data shown just above and below each line, respectively. Duos 1 to 10 are from replicate 1, and 11 to 20 are from replicate 2. In the first round of breeding, birds were released to mixed-sex aviaries on day 1 with nest boxes provided until day 45, and this was repeated in round two (with new sets of potential partners) lasting from day 51 to day 95. Not indicated is an additional two-week period in unisex groups just before the start of the second round. Some females divorced and re-paired after an initial pairing within one round (two green triangles within one round), most formed a single pair bond, and some others remained unpaired (missing triangle in the respective line).

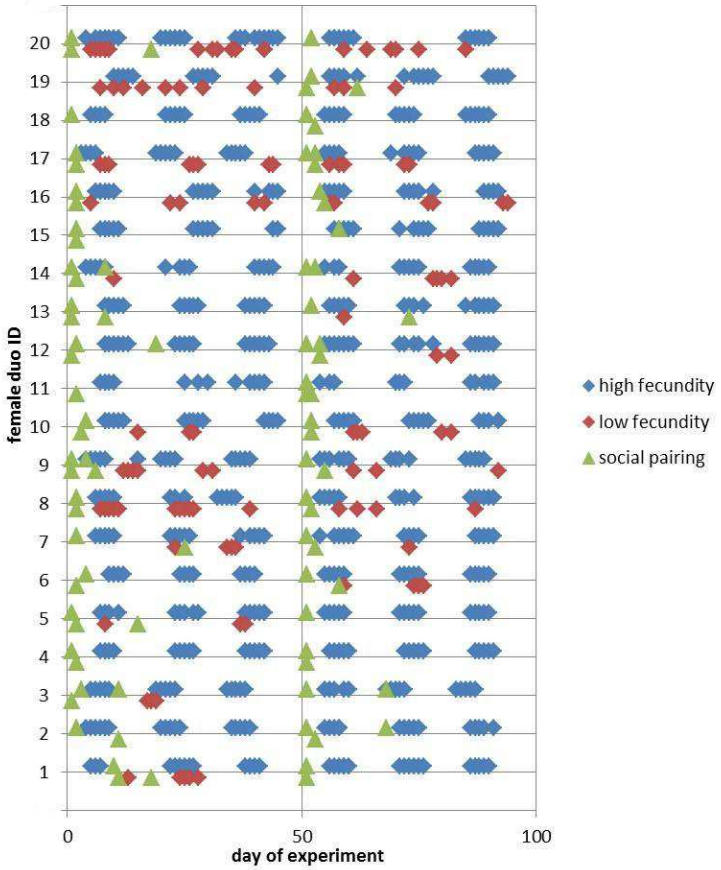
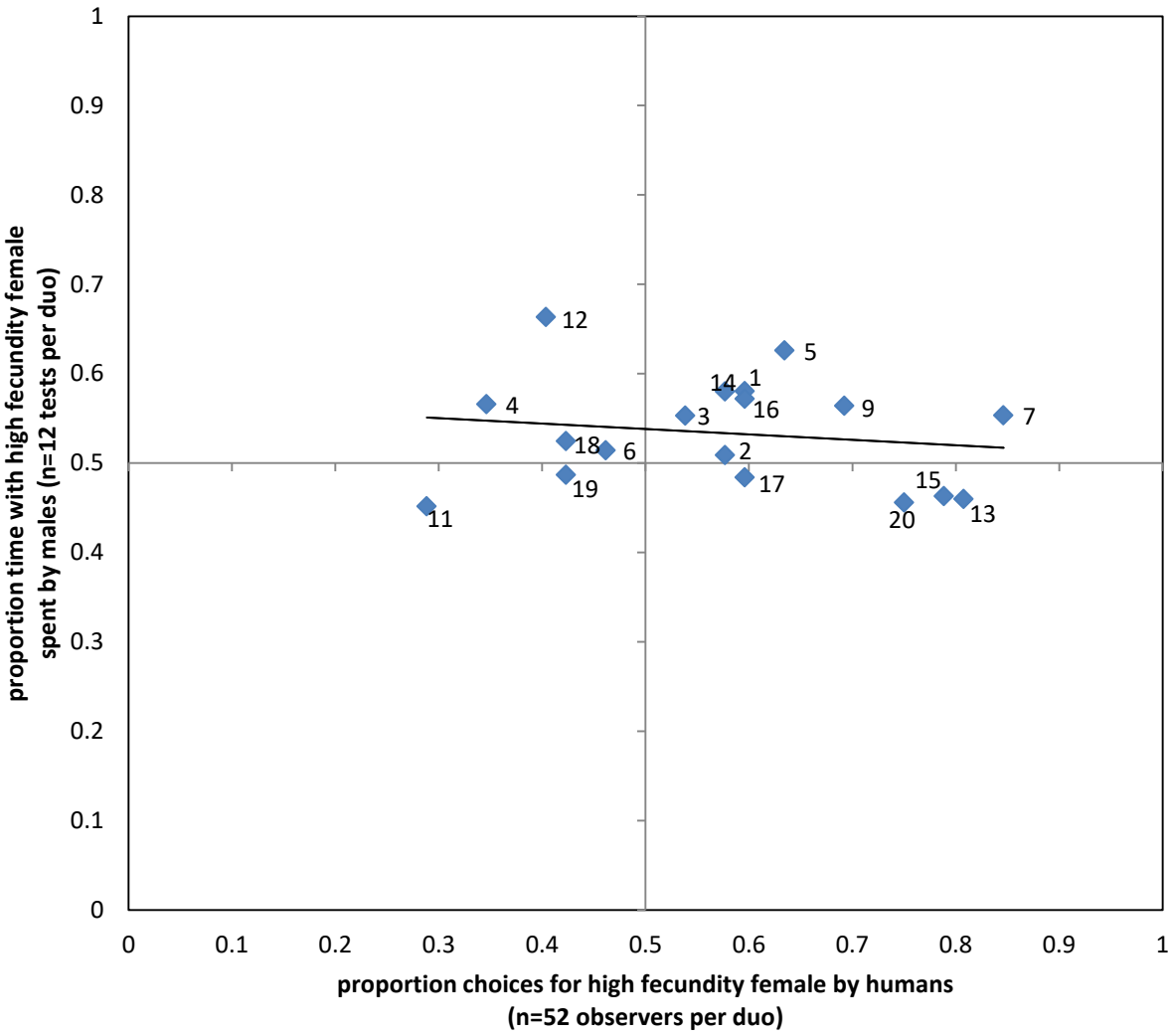


Figure S4. The relationship between human rating of the 18 female duos and the choices made by zebra finch males. The x-axis shows the proportion of 'correct' choices made by human observers (n = 52) for each female duo (blue diamonds labelled by duo ID as in Figure S3). The y-axis shows the proportion of time males associated with the high-fecundity female during choice-chamber tests (averaged for each duo across 6 males in a total of 12 tests). An ordinary least square regression line is shown.



Chapter 2: No mutual mate choice for quality in zebra finches: time to question a widely-held assumption

Short title: No mutual mate choice for quality in zebra finches

Abstract: Studies of mate choice typically assume that individuals will prefer high quality mates and select them based on condition-dependent indicator traits. In species where both sexes invest substantially in parental care, mutual mate choice is expected to result in assortative mating for quality. When assortment is not perfect, the lower quality pair members are expected to compensate by increased parental investment in order to secure their partner (positive differential allocation). This framework has been assumed to hold for monogamous model species like the zebra finch (*Taeniopygia guttata*), but little positive evidence has emerged, maybe because of the difficulty of defining individual quality. By combining multiple measures of causes (inbreeding, early nutrition) and consequences (ornaments, displays, fitness components) of variation in quality into a single principal component, we here show that quality variation can be quantified successfully and it indeed predicts individual pairing success, presumably because it reflects an individual's vigor or ability to invest in reproduction. Yet, despite high statistical power, we found no evidence for either assortative mating or for positive differential allocation. We suggest that zebra finch ornaments and displays are not sufficiently reliable for choosy individuals to obtain benefits from being selective about such traits that are greater than the costs of competition for the putative best partner. We call for unbiased quantification of preference strength and signal honesty and avoidance of selective reporting of significant results.

Published as: Daiping Wang, Wolfgang Forstmeier, Bart Kempenaers 2017: No mutual mate choice for quality in zebra finches: Time to question a widely held assumption. Evolution 71(11):2661-2676.

doi:10.1111/evo.13341

No mutual mate choice for quality in zebra finches: Time to question a widely held assumption

Daiping Wang,[1] Wolfgang Forstmeier,[1,2] and Bart Kempenaers[1]

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

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

Received April 30, 2017
Accepted August 23, 2017

Studies of mate choice typically assume that individuals prefer high quality mates and select them based on condition-dependent indicator traits. In species with biparental care, mutual mate choice is expected to result in assortative mating for quality. When assortment is not perfect, the lower quality pair members are expected to compensate by increased parental investment to secure their partner (positive differential allocation). This framework has been assumed to hold for monogamous species like the zebra finch (*Taeniopygia guttata*), but progress has been hampered by the difficulty to define individual quality. By combining multiple measures of causes (inbreeding, early nutrition) and consequences (ornaments, displays, fitness components) of variation in quality into a single principal component, we here show that quality variation can be quantified successfully. We further show that variation in quality indeed predicts individual pairing success, presumably because it reflects an individual's vigor or ability to invest in reproduction. However, despite high statistical power, we found no evidence for either assortative mating or for positive differential allocation. We suggest that zebra finch ornaments and displays are not sufficiently reliable for the benefits of choosiness to exceed the costs of competition for the putative best partner. To assess the generality of these findings unbiased quantification of signal honesty and preference strength is required, rather than selective reporting of significant results.

KEY WORDS: **Assortative mating, differential allocation hypothesis, fitness, mate choice, pairing status, pairing success, quality indicator.**

Most theories of mate choice predict that individuals should prefer high-quality over low-quality partners, because such a directional preference will typically be favored by selection (Kuijper et al. 2012). Here, "quality" refers to an individual's intrinsic propensity or ability to achieve fitness in an average environment (Wilson and Nussey 2010). Whenever potential partners vary in their intrinsic quality choosing individuals should aim for the highest quality partner they can secure (Andersson 1994). Between-individual variation in quality is expected to be ubiquitous (Wilson and Nussey 2010) due to both genetic effects (e.g., mutational load) and environmental factors (e.g., limited resources, changing selection pressures). A large body of mate choice literature has focused on documenting the extent to which ornaments or displays can function as honest signals of intrinsic quality (Hamilton and Zuk 1982; Nakagawa et al. 2007; Catchpole and Slater 2008; Dunn et al. 2010), and numerous studies have described directional mating preferences for such quality indicators (e.g., Andersson 1982; Welch et al. 1998; Reid et al. 2004; Pincemy et al. 2009; Doutrelant et al. 2012; Wells et al. 2015). It is therefore tempting to assume that directional mate choice preferences for quality indicators will be ubiquitous in the animal kingdom. Yet this assumption can be challenged for two reasons: a methodological one and an evolutionary one.

First, for each studied species, numerous potential quality indicators can be measured and preferences or choice outcomes can be quantified in many different ways. In empirical studies, this often leads to a considerable problem of multiple testing in combination with the risk of selective reporting of positive results (Forstmeier et al. 2016). This makes it difficult to judge how often the null hypothesis of no directional preference for

© 2017 The Author(s). *Evolution* © 2017 The Society for the Study of Evolution.

quality indicators might actually be true. Hence, we still need systematic assessments that ensure a comprehensive and unbiased reporting to calculate average effect sizes that include all "null findings."

Second, the condition-dependence of quality indicators is often limited (Cotton et al. 2004; Bolund et al. 2010b; Wang et al. 2017). Indicator trait values will then explain only a small amount of the true variance in quality, implying that selective individuals will–on average–obtain only small benefits from being choosy. Genetic variants leading to strong preferences for such weak indicator traits may then not spread to a high allele frequency in the population, because the obtained benefits might be smaller than the costs of being choosy. Such costs generally include time and sampling effort to identify a high-quality partner, but in socially monogamous species also include intensified competition for the most-ornamented partners once the preference has become widespread. Costs of competition may lead to selection favoring less-choosy or even nonchoosy individuals (Dechaume-Moncharmont et al. 2016).

Generally, we expect individuals to be choosy if they invest heavily in reproduction, for example in parental care. This is because the investment, for instance by females, makes these females unavailable for mating ("time out" of the mating pool, resulting in loss of potential mating chances, Kokko and Jennions 2008). In species where both males and females invest substantially in parental care, both sexes are expected to be choosy (Jones and Hunter 1993; Courtiol et al. 2016) unless the costs of choosiness are high or mates do not vary in quality (Kokko and Johnstone 2002). Under the assumption of preferences for high-quality individuals, mutual mate choice will then result in assortative mating by quality (Johnstone et al. 1996; Johnstone 1997; Bergstrom and Real 2000; Kokko et al. 2003; Hardling and Kokko 2005; Hooper and Miller 2008; Baldauf et al. 2009; Fawcett and Bleay 2009; Jiang et al. 2013). Such assortment simply arises from the fact that only high-quality individuals would be able to secure a high-quality partner, because the latter should reject any low-quality suitor. However, Burley (1988) suggested that a lower quality individual might also be able to obtain a higher quality partner, if the low-quality individual signals its readiness to invest relatively more in parental care, and subsequently would be able to secure the partner and maintain the pair bond by carrying out most of the workload (positive differential allocation). Such readiness to invest disproportionally in parental care represents another dimension of quality of one's partner. It is thus important to take differential allocation into account in studies of assortative mating for quality, because pairs that are not matched for quality may compensate the asymmetry in this way. If low-quality individuals are unable to invest disproportionally into parental care (i.e., unable to make themselves more attractive in this way), pairs that are poorly matched for quality are expected to be unstable. This should then lead to divorce and repairing until a level of assortment by quality is reached beyond that individuals would gain little from divorcing and trying to obtain a better partner.

Assortative mating for quality may arise not only in systems with mutual mate choice for quality, but also in monogamous systems where only one sex is choosy (Burley 1983; Fawcett and Johnstone 2003; Venner et al. 2010). If all members of the choosy sex would prefer high-quality individuals of the nonchoosy sex, intense competition among the members of the choosy sex for the best mate would imply that only the most competitive individual can secure the best mate (Fawcett and Johnstone 2003; Venner et al. 2010). Hence, to the extent that measures of individual quality also reflect competitive ability, we would expect assortment by quality also in this case.

Finally, assortative mating for quality can arise even in the absence of preferences for high-quality partners through indirect effects such as competition for high-quality habitat. Assortative mating for quality can then emerge through choice for characteristics that correlate with individual quality (e.g., Galipaud et al. 2013).

If no assortative mating for quality is detected—assuming sufficient statistical power—preferences for quality indicators may be absent or the ability to compare potential mates and switch to a better option may be limited (see Gimelfarb 1988a, 1988b).

Zebra finches (*Taeniopygia guttata*) are one of the most intensely studied organisms regarding mate choice (reviewed by: Collins and ten Cate 1996; Adkins-Regan 1998; ten Cate and Vos 1999; Riebel 2003; Adkins-Regan 2007; Riebel 2009; Griffith and Buchanan 2010; Hauber et al. 2010; Adkins-Regan 2011). Their predominant mating system is lifetime monogamy with both sexes investing about equally in parental care (Zann 1996). This high investment of both sexes is expected to favor choosiness in both males and females when searching for a (lifetime) partner (Kokko and Johnstone 2002). The species is abundant, breeds in dense colonies, and forms large flocks in the nonbreeding period, where new pair bonds can form long before reproduction (Zann 1996). This means that encounter rates of potential mates are presumably high, and hence the cost of being choosy during the period of pair formation should be low and should not hamper the evolution of choosiness (Johnstone 1997; Kokko and Johnstone 2002).

Research on zebra finches has stimulated the development of the differential allocation hypothesis (Burley 1988; Sheldon 2000; Ratikainen and Kokko 2010), and directional mutual mate choice for quality has been assumed for this species (Collins and ten Cate 1996; Riebel 2009, but see Forstmeier and Birkhead 2004; Ihle et al. 2015; Wang et al. 2017). However, despite the large body of zebra finch mate choice literature, only a few studies have reported significant assortative mating for a putative quality indicator (e.g., for natal brood size; Holveck and Riebel 2010).

Furthermore, there are no zebra finch studies that directly confirm the results of Burley (1988) to provide additional evidence for positive differential allocation.

One practical problem with studying assortative mating and differential allocation is the difficulty of quantifying variation in quality (Wilson and Nussey 2010; Bergeron et al. 2011; Lailvaux and Kasumovic 2011). Individual quality can be conceptualized as an individual's intrinsic propensity to achieve fitness (Wilson and Nussey 2010). However, fitness is also influenced by stochastic events that may act independently of the individual's phenotype (e.g., accidental destruction of a nest, sudden spell of bad weather, low food availability) and these events add noise to the relationship between achieved fitness and intrinsic individual quality. One cannot directly examine whether individuals pair assortatively for quality by looking at the correlation between male and female fitness, because the covariance is inflated by numerous shared effects that determine the fitness that the members of a pair achieve together and would equal one under strict genetic monogamy. This problem can be solved in experimental studies that take place under standardized conditions of captivity. Here, it is possible to measure fitness components achieved in one breeding round, and to use this as a predictor of who will pair with whom in a subsequent breeding round where individuals are exposed to a new set of potential partners. Despite statistical noise and despite the trade-off between current and future reproduction (Nilsson and Svensson 1996), such measures of fitness components can show considerable individual repeatability across pair bonds and breeding rounds (e.g., Bolund et al. 2011; Wang et al. 2017), reflecting the between-individual variation in intrinsic quality (Van Noordwijk and de Jong 1986). Thus, one can assess whether the pairs that form in a second breeding round pair assortatively with regard to the fitness they achieved in a first breeding round, and vice versa, whether the pairs that formed in the first round mated assortatively with regard to the fitness they achieved in the second round.

Individual phenotypic traits are often only weak predictors of true intrinsic quality. Hence, previous studies have used principal component analysis (PCA) to summarize all traits in a single quality score (Hamel et al. 2009; Moyes et al. 2009). Following the general recommendation by Wilson and Nussey (2010), this is the approach we adopted in this study, whereby we included a variety of quality indicators. First, intrinsic quality is known to be susceptible to both genetic and environmental stress, meaning that inbred birds and birds that faced harsh environmental conditions during early growth will typically achieve reduced fitness (Lindstrom 1999; Blount et al. 2003; Chapman et al. 2009; Bolund et al. 2010a; Forstmeier et al. 2012). Hence, we use between-individual variation in inbreeding coefficients (Forstmeier et al. 2012) and in early growth conditions (mass at 8 days of age (Bolund et al. 2010a), and natal brood size (Holveck and Riebel 2010)) as correlates of quality. Second, in zebra finches, a range of ornaments and displays seem susceptible to stressors such as inbreeding or malnutrition (e.g., beak color and plumage ornaments: Birkhead et al. 2006; Naguib and Nemitz 2007; Bolund et al. 2010a,b; song characteristics: Riebel 2009; Ritschard et al. 2010). If so, these ornaments or displays can be used by choosing individuals as phenotypic cues for identifying high-quality individuals. Third, we used measures of components of fitness (male siring success and female fecundity) that show considerable individual repeatability across pair bonds and seasons (e.g., Bolund et al. 2011; Wang et al. 2017), and reflect between-individual variation in intrinsic quality (Van Noordwijk and de Jong 1986). Based on PCA using all these measures, we considered an individual's score on the first principal component as a proxy for its intrinsic quality. We test how this score relates to pairing success and latency of pairing in communal breeding aviaries and whether pair bonds formed assortatively with regard to male and female scores. Finally, we test whether deviations from assortative mating (i.e., unequal quality of partners) predicts who takes the greater share in parental care.

Methods

STUDY POPULATION

We studied zebra finches from a domesticated population kept at the Max Planck Institute for Ornithology in Seewiesen, Germany (population # 18 in Forstmeier et al. 2007). Housing conditions, diet, and aviary specifications for breeding have been described in detail in the Supplementary File to (Wang et al. 2017). The population has been maintained at Seewiesen since 2004 (generations F1 to F4), and in 2009 we initiated the breeding of lines that were selected for high versus low breeding values for male courtship rate (two high lines, two unselected control lines, two low lines; see Mathot et al. 2013; Wang et al. 2017). The third generation of these six lines consists of a total of 343 females and 338 males. For this study, we used a subset of 219 females and 217 males (about equally representing the six lines) that participated in a breeding experiment (see below). All birds were color-banded for individual recognition.

VARIATION IN INBREEDING

Birds differed substantially in their inbreeding coefficient F (ranging from 0.005 to 0.299, see Table 1; calculated using Pedigree Viewer 6.4a, (Kinghorn and Kinghorn 2010), based on eight generations of pedigree data). Average F differed only slightly between the six lines (high 1: 0.12, high 2: 0.12, control 1: 0.12, control 2: 0.12, low 1: 0.11, low 2: 0.11; unpublished data). Inbreeding affected body size, ornaments, courtship display, and fitness measures in this population (see below Bolund et al. 2010a; Forstmeier et al. 2012).

Table 1. Summary statistics of all quality-related traits measured in male and female zebra finches (see Methods for details).

	Males			Females		
Trait	*N*	Mean ± SD	Range	*N*	Mean ± SD	Range
Inbreeding coefficient F	217	0.11 ± 0.06	0.005–0.299	219	0.12 ± 0.07	0.005–0.299
Natal brood size	217	3.39 ± 1.20	1–6	219	3.25 ± 1.23	1–6
Mass day 8 (g)	217	7.47 ± 1.30	3.3–10.2	219	7.37 ± 1.53	3.1–11.6
Mass day 100 (g)	217	14.95 ± 1.19	12.0–20.9	219	15.42 ± 1.43	12.5–20.0
Beak color (score)	217	3.71 ± 0.36	2.60–4.40	219	2.63 ± 0.32	1.5–3.6
Courtship rate	217	3.26 ± 1.82	0.00–8.14			
Cheek patch size (mm^2)	200	124.1 ± 12.6	94.3–158.3			
Amplitude (db)	191	32.25 ± 2.08	26.50–38.21			
Repertoire size	191	4.67 ± 1.25	2–8			
Eggs sired round 1	216	7.52 ± 6.25	0–26			
Eggs sired round 2	216	8.00 ± 6.74	0–31			
Fecundity round 1				216	9.22 ± 4.86	0–22
Fecundity round 2				216	9.21 ± 5.26	0–18

MEASUREMENTS OF EARLY REARING CONDITIONS

When nestlings were 8 days old, which is roughly the time of maximal growth, we measured body mass to the nearest 0.1 g using an electronic scale. This measure is highly variable (Table 1), has a low heritability and is known to primarily reflect conditions during early growth (Bolund et al. 2010b).

For each individual, we also defined its natal brood size as the number of chicks that reached 8 days of age in the nest it had fledged from (Table 1). We included natal brood size in our test for assortative mating because several studies on zebra finches suggested that natal brood size affects quality, with birds originating from smaller broods being higher quality individuals (DeKogel and Prijs 1996; Tschirren et al. 2009; Holveck and Riebel 2010), but see (Kriengwatana et al. 2016). Estimates of heritability of natal brood size in our population are close to zero (unpublished data), presumably because it depends on nonheritable parental effects (maternal fecundity, paternal fertility, embryo mortality, parental effort).

MEASUREMENT OF ADULT BODY SIZE, ORNAMENTS, AND DISPLAY TRAITS

When individuals were about 100–120 days old, that after reaching sexual maturity, we measured body mass to the nearest 0.1 g using an electronic scale. This measure reflects final adult size and condition (Table 1). On the same occasion, we also scored beak color according to the Munsell color chip system (Forstmeier and Birkhead 2004). Beak color is a condition-dependent trait affected by early growth conditions and inbreeding (Bolund et al. 2010a; Bolund et al. 2010b), with redder beaks indicating better condition in both sexes (Table 1).

We measured the size of each male's orange cheek patches between July and September 2015 after the birds had been breeding in aviaries (see measurement of "fitness components" below). We assumed that this trait does not change with age in a condition-dependent manner, that is that it was not influenced by the previous breeding experience. We took standardized photographs of the right and left cheek patch of each individual, as previously described (Bolund et al. 2010a). From the photographs, we measured the size of each patch using manual delineation of boundaries in the Image J software (Abramoff et al. 2004). The size measurements from the two photos were highly repeatable (left vs right cheek patch: Pearson $r = 0.68$, $n = 200$ males). For further analysis, we used the average value for each male (Table 1). Previous studies present evidence that cheek patch size is a condition-dependent indicator of male quality (Naguib and Nemitz 2007) but see (Bolund et al. 2010a) and a target of female choice (Roberts et al. 2007; Tschirren et al. 2012).

We measured male courtship rate, defined as the number of seconds of song that males directed towards unfamiliar females in a five-minute encounter. Each male participated in four five-minute trials (two at the age of 104–140 days and two at 200–228 days). During each trial we measured the total duration (in seconds) of song toward an unfamiliar female (mean: 16 s, range: 0–93 s, $n = 868$ encounters). Courtship duration was first square-root transformed to approach a normal distribution and then averaged for each male across the four trials (Table 1). Previous studies suggest that courtship rate is a quality-indicator (Houtman 1992; Bolund et al. 2010a). The trait is partly heritable (Houtman 1992; Forstmeier et al. 2011) and after three generations of directional selection on breeding values for courtship rate, males from high lines differed from those of low lines by more than two phenotypic standard deviations (Cohen's d = 2.36, 95% CI: 2.07–2.64; our unpublished data). Given that

individuals from the six lines show about equal fitness in communal breeding aviaries (details of communal breeding set up see below; each aviary contained one male and one female from each selection line), genetic differences in courtship rate may not be indicative of male quality, but the trait is condition-dependent in the sense that it is affected by inbreeding (Bolund et al. 2010a).

To examine specific song characteristics, we recorded the song of all males immediately after they participated in the breeding experiment, assuming that song measures did not change in relation to an individual's previous breeding experience. Methods of song recording and analysis were similar to those reported in (Forstmeier et al. 2009). In brief, to elicit courtship song, each male was placed together with an unfamiliar female in a metal wire cage equipped with three perches and containing food and water. The cage was placed within one of two identical sound-attenuated chambers. We mounted a Behringer condenser microphone (C-2, Behringer International GmbH, Willich, Germany) at a 45° angle between the ceiling and the side wall of the chamber, such that the distance to each perch was approximately 35 cm. The microphone was connected to a PR8E amplifier (SM Pro Audio, Melbourne, Australia) from which we recorded directly through a M-Audio Delta 44 sound card (AVID Technology GmbH, Hallbergmoos, Germany) onto the hard drive of a computer. We used Sound Analysis Pro version SAP 2011 with a sampling rate of 44 kHz and 16 bit amplitude resolution (Tchernichovski et al. 2004).

From the song recordings we selected, for each male, two representative high-quality motifs (a motif is a more or less stereotypical, repeated part of a male's song; Forstmeier et al. 2009). Syllables within a motif were automatically delineated by setting a fixed amplitude and entropy threshold in the SAP software. As previously described (Forstmeier et al. 2009) we obtained two song characteristics: repertoire size (number of different syllables) and song amplitude (sound volume). Several characteristics of male song have been suggested as indicators of male quality (Riebel 2009), but we selected these two parameters because (1) amplitude was the only trait that was significantly positively related to male siring success in an earlier study on the same population (estimate $\pm$ SE from mixed effect model: 0.35 ± 0.10, $t = 3.50$, $P = 0.0005$; our unpublished data), and (2) repertoire size is a widely studied song trait related to aspects of male quality (e.g., Spencer et al. 2003, 2005; Boogert et al. 2008; Vyas et al. 2008; Soma and Garamszegi 2011; Woodgate et al. 2011, 2012).

All measurements described above were taken blind with respect to the other characteristics of an individual (except for mass at day 100, which was taken after scoring beak color, but observer bias is unlikely to affect mass measurements with an electronic scale).

ASSESSING PAIR BOND FORMATION AND MEASURING FITNESS COMPONENTS

We placed the birds in aviaries for breeding between January 2014 and May 2015. Males and females from each of the six selection lines were first randomly assigned to one of four cohorts each comprising 54 males and 54 females (216 of each sex in total). Birds in each cohort were then assigned to one of nine aviaries, such that (1) each aviary contained one male and one female from each selection line (nine aviaries x six lines) and (2) all birds within an aviary were unfamiliar with each other. In 44 cases, we did not have enough males or females from a certain selection line, so we instead used birds from the same line type (high, control, or low).

Birds were allowed to freely choose a mate within their aviary and they produced up to three clutches within a period of seven weeks (referred to as "one breeding round"; nest boxes were provided from day 1 to day 45). During daily nest checks, all eggs laid were immediately replaced by plastic eggs and placed in an incubator for four days, that is until an embryo had formed for later parentage assignment. Clutches consisting of plastic eggs were removed after 10 days of incubation to allow the female to lay the next clutch. On day 49, individuals were separated by sex into different rooms for a two-week period, after which we initiated an identical second round of breeding, but with a different set of potential partners (by putting the six males of one aviary into the next room). In this second round only, females were familiar with on average one quarter of the males because they grew up with them in the same peer group (there were four mixed-sex peer groups holding the juveniles from 35 to 100–120 days of age). We did not have enough birds in different peer groups to avoid this. Across the four cohorts, one male and three females that died during the first breeding round were replaced by an individual from the same line in the second round, leading to a total of 217 males and 219 females participating in the experiment.

During each breeding round of a cohort, we carried out daily observations on individuals allopreening, sitting in body contact, and visiting a nestbox together, which reflects pair bonding as previously described (Wang et al. 2017). Observations lasted approximately 30 min (total across the nine aviaries) and were carried out approximately 120 times per breeding round. We decided that a pair bond had been formed when at least eight records of pair bonding behavior had been recorded for that pair. Some individuals engaged in multiple pair bonds, either sequentially (considered as monogamous) or simultaneously (polygamous). Because females were less polygamous than males (see below) we decided to focus on the female perspective to define the start and end of a pair bond. We defined the start of one pair bond as the time when the female restricted her pair bonding behavior to a single male (or, in rare cases of clear polyandry, to two males). The end of a pair bond was defined by either the time when the

Table 2. Principal component analysis (PCA) of traits reflecting male and female quality (Table 1).

	Males			Females		
Trait	M PCA mean	M PCA round1	M PCA round2	F PCA mean	F PCA round1	F PCA round2
Natal brood size*	0.47	0.48	0.45	0.00	0.01	–0.02
Mass day 8	0.37	0.33	0.41	0.78	0.80	0.77
Mass day 100	0.50	0.46	0.56	0.78	0.78	0.79
Inbreeding_coefficient F*	–0.37	–0.34	–0.32	–0.22	–0.19	–0.24
Beak color	0.14	0.12	0.26	0.45	0.46	0.45
Cheek patch size	0.54	0.57	0.58			
Courtship rate	0.22	0.23	0.26			
Repertoire size	0.12	0.06	0.10			
Amplitude	0.47	0.54	0.39			
Mean siring success	0.55					
Siring success round 1		0.56				
Siring success round 2			0.41			
Mean fecundity				0.48		
Fecundity round 1					0.44	
Fecundity round 2						0.47
Total variance	0.17	0.17	0.16	0.28	0.28	0.28
Eigenvalue observed	1.66	1.67	1.60	1.71	1.69	1.70
Eigenvalue simulated	1.35	1.35	1.35	1.23	1.23	1.23
(95% CI)	(1.25–1.47)	(1.25–1.47)	(1.25–1.48)	(1.13–1.35)	(1.14–1.34)	(1.14–1.34)

For males, each PCA contains the same nine phenotypic traits and one of three measures of siring success. For females, five phenotypic traits and one of three measures of fecundity are included. Shown are the loadings (correlation coefficients) of each trait on the first principal component, the proportion of the total variance explained by the first principal component (Total variance), and its eigenvalue (observed). This value is compared against an average eigenvalue (and its 95% confidence interval) based on 10,000 simulation runs where trait values were randomized among individuals. Traits for which a negative loading on the first principal component was expected a priori are indicated with an asterisk. All other traits were expected to load positively.

female showed pair bonding behavior exclusively to a different male (start of a new pair bond) or the last observation of pair bonding behavior. Across the four cohorts and the two breeding rounds we identified a total of 423 pair bonds. Of these, 342 bonds were classified as monogamous and 292 of them lasted to the end of the breeding round (the remainder led to divorce or ended due to the death of a partner). The remaining 84 relationships were polygamous, including 29 polygynous males with 29 primary and 33 secondary females (somewhat arbitrarily ranked by the amount of pair bond behavior observed and sometimes by order of pairing), and 11 polyandrous females with 11 primary and 11 secondary males.

In total, 4041 eggs were laid, but from 685 of them no DNA was obtained (14 tiny eggs without yolk, 24 broken eggs, 632 apparently infertile eggs, 15 cases of lost samples or DNA concentration too low). The 685 untyped eggs were either assigned to the mothers who attended the nest (625 eggs, assuming no egg dumping) or remained unassigned if the nest was unattended (60 eggs). The remaining 3356 eggs were unambiguously assigned to parents using 15 microsatellite markers (see Wang et al. 2017), but four eggs were only assigned to their mother (due to parthenogenesis, mosaicism, or siring by sperm from the previous experimental round). Thus, overall, 3352 eggs were allocated to their genetic father and 3981 eggs were allocated to their mother.

As measures of fitness components, we counted for each male the total number of eggs sired in the first and in the second breeding round ("eggs sired," Table 1). For females we counted the total number of eggs laid in each round ("fecundity," Table 1). We also calculated for each bird the average number of eggs sired or laid over the two breeding rounds. Due to four birds dying during the first round of breeding (see above), fitness measurements were available from both breeding rounds and for both pair members for 417 (rather than 423) pairs.

CALCULATION OF INDIVIDUAL QUALITY SCORES

For the principal component analysis of male quality, we included ten traits, as shown in Table 2. For the trait "siring success" there were three measures for each male: "siring success round 1," "siring success round 2," and "mean siring success." We conducted a different PCA on each of them in combination with the other nine traits ("M PCA round1," "M PCA round2," "M PCA mean"). The first two allowed us to extract male quality scores (the first principal component) for each individual based on fitness data from the first or second round only, which were used to avoid covariance

between the PCA scores of the members of a pair that arises from the fitness that they achieved when they actually bred together.

For PCA of female quality, six traits were included (Table 2). As for male siring success, we also conducted a different PCA for each of three fecundity measures ("F PCA round1," "F PCA round2," "F PCA mean").

Principal components were calculated using the "prcomp" function in R (R Core Team 2015) based on the correlation matrix while replacing a few missing values (see sample sizes in Table 1) by mean values. To evaluate the magnitude of the Eigenvalue of the first principal component, we simulated 10,000 PCAs where trait values were randomized among individuals (without replacement), that is a situation where traits are not systematically correlated with each other. We then compared the Eigenvalue of the first principal component of the true data to the 95% CI of the simulated Eigenvalues.

RELATING INDIVIDUAL QUALITY SCORES TO PAIRING SUCCESS

First, we quantified pairing success within each breeding round as the sum of the number of days an individual was paired to any opposite-sex individual (across the 49 days breeding period). For instance, a male that was paired monogamously from day 2 to day 49 of the breeding period had a pairing success of 48 days, while a polygynous male with a primary female from day 1 to day 49 and a secondary female from day 10 to day 35 had a pairing success of $49 + 26 = 75$ days. To test whether high quality individuals were more successful at forming and keeping pair bonds, we regressed pairing success in a given round over the individual's principal component score from the other breeding round. Thus, measurements of the total siring success of males and of fecundity of females (measured across both rounds of the experiment) were replaced with measurements from either the first or the second round. In this way, principal components can be used as predictors of success in a given round without being influenced by the success obtained in that round. Because the dependent variable (days paired) was clearly not normally distributed we derived *P*-values for regression slopes from randomization tests. The dependent variable was randomized 100,000 times (without replacement) among the available predictor values, yielding 100,000 regression slopes. The proportion of randomly generated slopes that was steeper than the observed slope (in absolute terms, i.e. two-tailed test) was taken as the *P*-value.

Some models of mate choice predict that high-quality individuals get to choose their partner first (Bergstrom and Real 2000). Thus, we also tested whether PC scores were related to the latency to pair (in days), focusing only on the first two weeks of each experimental round, which is the period when approximately 80% of the birds had paired (birds that paired later were excluded in this analysis).

Some models of mate choice in monogamous systems emphasize the temporal dynamics of individuals pairing up such that later-pairing individuals have fewer choices (Gimelfarb 1988b). These models predict changes in the degree of assortative mating over time (as more pair bonds are established). We therefore show how the average quality (PC scores) of mating males and females as well as the level of assortment (Pearson correlation coefficient between PC scores of pair members) change with pairing order within aviaries. To this end, all pair bonds were ranked within aviaries by timing of pair formation, and statistics were calculated within each rank (rather than cumulatively as done in Gimelfarb (1988b)) across the 72 aviaries.

Finally, we also tested whether individuals with high PC scores were more likely to be polygamous. Specifically, we tested whether PC scores differed between unpaired and polygynous males, or between unpaired, and polyandrous females. To avoid pseudo-replication, separate tests were carried out for the first and second breeding round.

TEST OF ASSORTATIVE MATING FOR MEASURES OF QUALITY

To quantify the strength of assortative mating at the population level, we calculated Pearson's correlation coefficients between homologous traits of the partners across all 423 pair bonds that were observed. This population-wide measure is analogous to measurements of assortment based on studies in the wild, where the range of potential partners available to each individual is typically unknown. Assortment was calculated for all homologous traits (inbreeding coefficient, natal brood size, mass at day 8, mass at day 100, and beak color), as well as for male siring success versus female fecundity (both from the other breeding round), and for male and female PC scores (also both from the other round).

We also calculated the strength of assortment at the aviary level, because–unlike in the wild–we have information about all available potential partners (all individuals placed together in one of the 72 experimental aviaries). To this end, we calculated 72 Pearson's correlation coefficients for each of the seven homologous traits mentioned above. In aviaries with only few pair bonds, some variables (e.g., natal brood size) did not vary in one sex, so the number of calculated correlation coefficients varied between 69 and 72. Correlation coefficients were subjected to Fisher's z transformation to approach normality, then averaged across all aviaries, and tested against zero using a two-tailed one sample *t*-test.

To investigate whether the strength of assortative mating for quality (PC scores) varies with pairing status, we calculated population-wide Pearson's correlation coefficients for different subsets of pair bonds (all 342 monogamous pair bonds, all 292 monogamous pair bonds that lasted until the last day of a breeding round, all 84 polygamous relationships).

TEST FOR DIFFERENTIAL ALLOCATION

During the daily observations and nest checks, we also recorded which individual of a pair was inside the nest (presumably incubating the clutch; birds were rarely inside the nest when there were no eggs or only cold eggs). To test whether the lower quality individual of a pair carried out a larger proportion of the parental care, we focused on the subset of 292 monogamous pair bonds that lasted to the end of the breeding round (other subsets were not examined). Of these pairs, 283 had been recorded inside the nest in a total of 3431 instances. For these pairs we modelled the proportion of female incubation (using the "cbind" function in R on female and male counts of presence inside the nest with a binomial error distribution) as the response variable. Female ID, male ID, and pair ID were added as random effects (the latter to control for overdispersion of counts within pairs). As the fixed effect of interest, we fitted the estimated difference in female versus male quality (female PC score minus male PC score for each pair, using PC scores from the other breeding round).

STATISTICAL ANALYSES

All simulations and statistical analyses were done using R 3.2.3 (R Core Team 2015). For PCA we used the "prcomp" function, for Pearson's correlation we used the "cor.test" function, and for mixed-effect models we used the lme4 package (Bates et al. 2015). To allow assessment of the repeatability of our measurements, we report Pearson correlation coefficients for traits that were measured twice (first vs second measurement). For all mixed-effect models, we report the proportion of variance explained by the random effects (relative to the sum of all random effects plus residual) after accounting for the fixed effects.

Results

Females that laid many eggs in the first breeding round also laid many eggs in the second round ($r = 0.63$, $n = 213$, $P < 0.0001$; Table S1), and males that sired many eggs in the first round also did so in the second round ($r = 0.46$, $n = 215$, $P < 0.0001$; Table S2). This indicates substantial between-individual variation in reproductive investment or in the ability to achieve fitness (independent of the partner and social environment). However, other potential indicators of individual quality showed at best moderate correlations with either female fecundity (averaged over both breeding rounds; largest $r = 0.20$ for mass at day 8 and mass at day 100, $n = 219$, $P = 0.003$; Table S1) or with male siring success (largest $r = –0.27$ for the inbreeding coefficient, $n = 216$, $P < 0.0001$; Table S2).

SUMMARIZING INDIVIDUAL QUALITY BY PCA

Principal component analysis of the 10 indicators of male quality resulted in a first PC ("M PCA mean") that explained only 17% of the total variance in the data (Table 2). However, its eigenvalue of 1.66 was still notably higher than the random expectation of 1.35 (95% CI: 1.25–1.47, Table 2). Furthermore, nine of the ten putative indicators of quality showed loadings in the expected direction (Table 2). Only the loading of natal brood size went against the expectation, suggesting that males from larger broods showed phenotypic traits associated with higher (rather than lower) quality. Of all quality indicators, the siring success that males achieved across both breeding rounds showed the strongest association with the first principal component ($r = 0.55$).

In females, the first principal component of the six quality indicators ("F PCA mean") explained 28% of the total variance (Table 2). Its eigenvalue of 1.71 was also higher than that expected under randomness (mean: 1.23, 95%CI: 1.13–1.35, Table 2). All loadings were in the expected direction, but natal brood size was not correlated with the first principal component (Table 2). Of all female quality indicators, measures of female body mass showed the strongest positive associations with the first principal component ($r = 0.78$), followed by female fecundity ($r = 0.48$).

For both sexes, the principal components from each breeding round separately showed nearly identical loadings (Table 2).

PC QUALITY SCORES AS PREDICTORS OF INDIVIDUAL PAIRING SUCCESS

Figure 1 illustrates how pairing success of individual males and females in a given breeding round was predicted by the principal component scores summarizing individual quality measurements (taken independently of that breeding round). Randomization tests showed that all four regression lines were significantly steeper than expected by chance (Fig. 1). Despite violation of the normality assumption, *P*-values were identical to those obtained from simple linear models assuming a Gaussian error distribution (details not shown).

Regression slopes of pairing success over the PC score were steeper in males (linear model estimates $\pm$ SE, round 1: 6.4 ± 1.5, round 2: 8.1 ± 1.2) than in females (round 1: 3.3 ± 1.2, round 2: 3.9 ± 1.2).

Across both breeding rounds we identified a total of 423 pair bonds. In 342 cases, these were strictly monogamous, that is none of the partners maintained an additional relationship during the same time period. However, some monogamous relationships were only of short duration, leaving 292 monogamous pair bonds that lasted until the end of the breeding round. A total of 84 relationships were polygamous. High-quality individuals, as identified through PCA, were significantly more likely to become polygamous, while low-quality individuals often remained unpaired (Fig. S1).

In males, there was a weak tendency for high-quality individuals to pair earlier (counting only the first pair bond of each

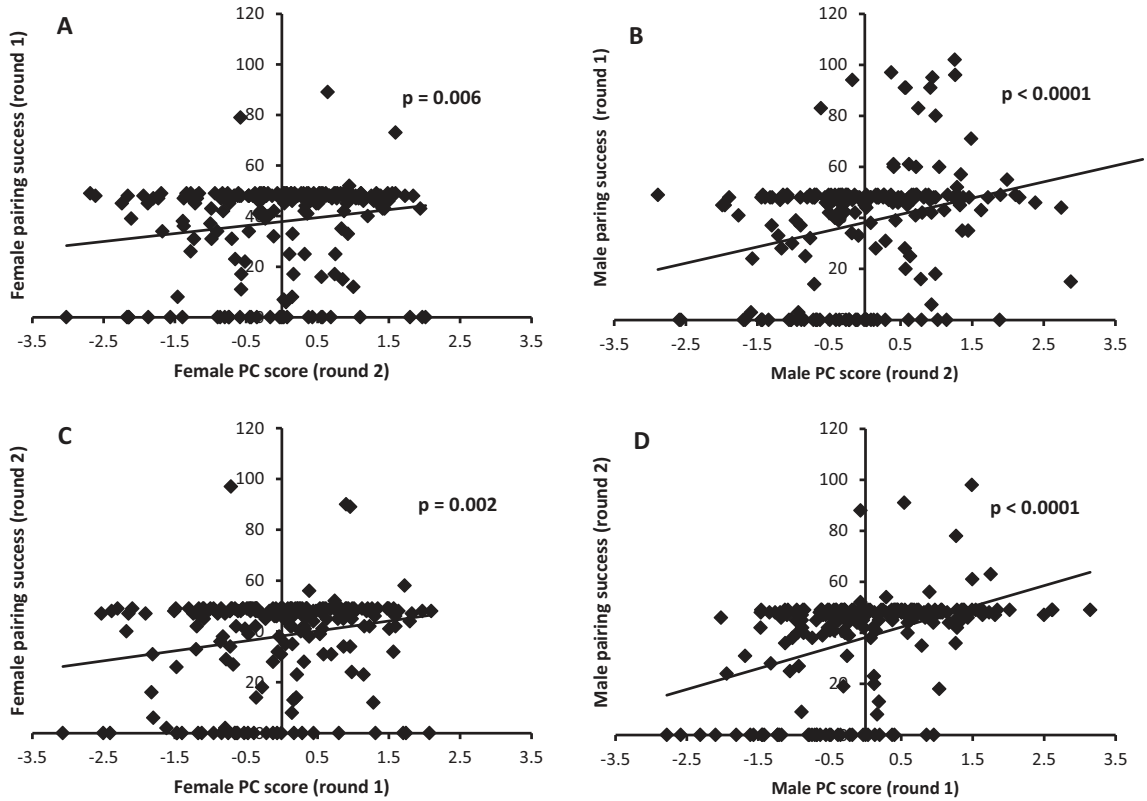


Figure 1. **Pairing success of individuals as a function of their quality (scored as the first principal component reflecting the quality indicators given in Table 2). Shown is the pairing success of females and males in the first breeding round (panels A and B), and in the second round (C and D). The *y*-axes show the cumulative number of days that an individual was paired during a 49-day breeding round (to any opposite-sex individual). Most birds paired to a single partner shortly after the start of a breeding round and remained paired until the end (most values in the 47–49 days range); birds with values > 49 days showed pair bonding behavior with multiple partners simultaneously (polyandry in females, polygyny in males; zero values represent birds that failed to form a pair bond. Note that the principal component scores (*x*-axes) were calculated from data obtained during the other breeding round (i.e., with a different partner). Ordinary least-squares regression lines are shown together with *P*-values based on randomization tests (100,000 simulations, see Methods).**

individual and including only individuals that paired within the first two weeks: Spearman-rank correlation of latency in round 1 vs M PCA round2: $r_s = –0.11$, $n = 164$, $P = 0.16$; latency in round 2 vs M PCA round1: $r_s = –0.16$, $n = 172$, $P = 0.04$). In females, these correlations were even weaker (latency in round 1 vs F PCA round2: $r_s = –0.05$, $n = 176$, $P = 0.55$; latency in round 2 vs F PCA round1: $r_s = –0.03$, $n = 175$, $P = 0.68$). Figure S3 shows how the quality (PC scores) of paired individuals and the level of assortative mating changed with pairing order (rank within aviaries).

ASSORTATIVE MATING FOR MEASURES OF QUALITY

For all 423 pair bonds, we found no evidence for positive assortative mating for quality as assessed through PCA (Fig. 2; $r = –0.05$, $P = 0.29$). Across the 211 pairs that formed in the first round, PC scores of pair members (calculated with data from the second round) were not significantly correlated ($r = –0.03$, $P = 0.67$; Fig. 2A). Likewise, the 212 pairs that formed in the second round showed no significant correlation in their PC scores (based on the first round; $r = –0.08$, $P = 0.27$; Fig. 2B). The population-wide pattern was not different from the analysis at the within-aviary level (average $r = –0.02$, $n = 72$ aviaries, $P = 0.79$; Table 3).

When running these analyses separately for each of the underlying traits that can be measured in both sexes, either at the population level or per aviary, 10 out of 14 correlation coefficients were negative and 12 out of 14 were not significantly different from zero (without adjustment for multiple testing; Table 3). The significant negative correlations were for natal brood size at the population level ($r = –0.10$, $P = 0.04$) and for body mass at day 100 at the within-aviary level (average $r = –0.14$, $P = 0.02$; Table 3).

Considering only certain subsets of pairs did not alter the conclusions. When excluding the 84 polygamous relationships, the remaining 342 monogamous pairs also showed no assortative

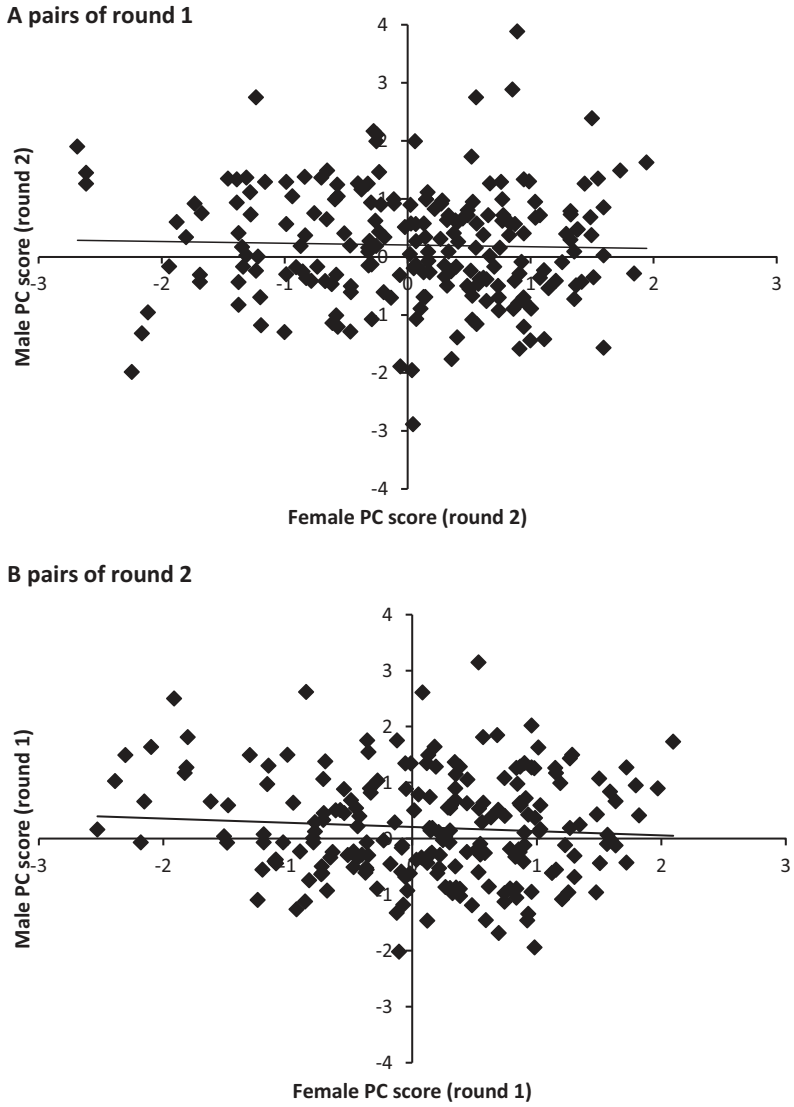


Figure 2. **Principal component scores reflecting individual quality of males and females that formed a pair bond during the first (A) and second (B) breeding round. Each dot represents a pair bond (n = 211 in (A) and n = 212 in (B)). Note that PC scores are calculated from data of the other breeding round. Ordinary least squares regression lines indicate the absence of assortative mating.**

mating at the population level for the principal component score reflecting quality ($r = –0.04$, $P = 0.52$; Table 4). This was also true for the subset of 292 monogamous pairs whose pair bond lasted until the end of the respective breeding round ($r = –0.004$, $P = 0.95$; Table 4).

Whether a monogamous pair bond lasted to the end of a breeding round ($n = 292$) or was split up (divorce, $n = 42$) did not depend on the quality difference between the partners, but the trend was in the expected direction (logistic regression: divorce predicted by the absolute difference in PC scores of the pair members, slope $\beta = 0.21 \pm 0.18$, $t = 1.2$, $P = 0.24$). Across these 42 cases of divorce, in 16 cases both partners repaired with other individuals, in 14 cases only the higher quality member repaired, in five cases only the lower quality member repaired, and in seven cases neither of the former pair members repaired.

Polygynous males were mostly of above-average quality (Fig. S1, Fig. S2), but their primary or secondary female partners were of average quality (Fig. S2). Likewise, polyandrous females were often of higher than average quality, but they were mated to average-quality males (Fig. S2).

TEST FOR DIFFERENTIAL ALLOCATION

The parental investment by the female relative to that of her male partner (proportion of incubation; measured as presence on the nest, see Methods) was independent of the difference in quality between the pair members as measured by their PC scores (female score minus male score; $\beta = 0.07$, $P = 0.18$; Table 5). The observed trend was in the opposite direction as expected, suggesting that, if anything, high-quality females paired to

Table 3. Test for assortative mating for various quality traits of males and females (Table 1) calculated at the population-wide and the within-aviary level.

	Population-wide					Within aviary				
Trait	Pearson r	*t*	df	*P*	*n*	Pearson *r* (mean)	*t*	df	*P*	*n*
Natal brood size	–0.1	–2.04	421	0.04	423	–0.04	–0.73	68	0.47	69
Mass day 8	–0.08	–1.68	421	0.09	423	–0.11	–1.53	71	0.13	72
Mass day 100	–0.08	–1.74	421	0.08	423	–0.14	–2.43	71	0.02	72
Inbreeding coefficient F	–0.01	–0.2	421	0.84	423	–0.07	–0.87	70	0.39	71
Beak color	0.05	0.93	421	0.36	423	0.05	0.66	69	0.51	70
Fitness	0.01	0.2	415	0.84	417	0.02	0.3	70	0.76	71
PCA score	–0.05	–1.06	421	0.29	423	–0.02	–0.27	71	0.79	72

At the population level, sample size (*n*) indicates the number of pairs. At the aviary level, the sample size indicates the number of aviaries for which a Pearson correlation coefficient was calculated; shown are the mean correlation coefficients for each trait. "Fitness" refers to the number of eggs laid (female fecundity) or fertilized (male siring success). PCA score refers to the estimate of overall quality (Table 2) based on the other breeding round, that is when not paired to the focal partner.

Table 4. Test for assortative mating for male versus female estimates of quality (PCA scores, Table 2) across pairs of various status categories (all pair bonds observed, all exclusively monogamous relationships, the subset of monogamous relationships that lasted until the end of a breeding round, and all polygamous relationships).

Pairing status	Pearson *r*	*t*	df	*P*	*n*
All pairs	–0.05	–1.06	421	0.29	423
Monogamous all	–0.04	–0.65	340	0.52	342
Monogamous lasting	–0.004	–0.07	290	0.95	292
Polygamous	–0.10	–0.85	79	0.40	84
Male polygyny	0.05	0.39	60	0.69	62
Female polyandry	0.07	0.29	20	0.77	22

Table 5. Results from a linear mixed-effect model testing the differential-allocation hypothesis based on observations of the proportion of female incubation in 283 out of 292 lasting monogamous pairs that initiated breeding.

Variable	Estimate V or β ± SE	z	*P*
Random effects			
Male ID ($n = 185$)	0.298		
Female ID ($n = 181$)	0.330		
Pair ID ($n = 283$)	0.083		
Fixed effects			
Intercept	0.108 ± 0.08	1.43	0.15
Quality difference	0.069 ± 0.05	1.34	0.18

The dependent variable is the relative count of female versus male nest visit records (using the "cbind" function in R) within each pair (average number of records per pair: female 6.5, male 5.6). The random effects male and female identity (ID) reflect the individual repeatability across different partners (variance component V), while the random effect pair ID controls for overdispersion in the binomial counts. The positive intercept (on the logit scale) reflects a greater effort by females than by males. The predictor of interest (quality difference) is the difference in estimated quality between the partners (female PCA score minus male PCA score). The sign of the parameter estimate ($\beta > 0$) is consistent with negative rather than positive differential allocation (see text for details).

low-quality males tended to do a greater share of incubation ("negative differential allocation").

Discussion

Summarizing between-individual variation in quality by means of principal component analysis (as recommended by Wilson and Nussey 2010) appears to have been successful: nearly all predictors of quality showed loadings in the expected direction, and the resulting PCA scores predicted individual pairing success. However, there was no evidence for positive assortative mating either by PC scores or by any of the underlying traits. Thus, this study neither is in line with the predictions of models stating that mutual mate choice for quality leads to assortative pairing (Johnstone et al. 1996; Johnstone 1997; Bergstrom and Real 2000; Kokko et al. 2003; Hardling and Kokko 2005; Hooper and Miller 2008; Baldauf et al. 2009; Fawcett and Bleay 2009), nor does it support models predicting that choosiness of one sex in combination with competition for the highest quality mates leads to assortment by quality (Burley 1983; Fawcett and Johnstone 2003; Venner et al. 2010). PC scores were only weakly related to pairing order in males (average $r_s = -0.13$), and in females (average $r_s = -0.04$), providing little support for the idea that the highest quality individuals get to choose their partner first (Bergstrom and Real 2000). Furthermore, there was no support for differential allocation based on quality differences (Burley 1988), suggesting

that lower quality individuals did not provide more parental care to secure their higher quality partner.

Our results may be counterintuitive and seem at odds with the dominant view from the avian mate-choice literature. We propose that our "null finding" is not due to inappropriate methods or lack of statistical power, but suggests that individual quality is not the target of mate choice in zebra finches. Our results are consistent with the observation that zebra finches show remarkably little between-individual agreement regarding the attractiveness of opposite-sex individuals, both in females (Forstmeier and Birkhead 2004; Ihle et al. 2013, 2015) and in males (Wang et al. 2017). Such low levels of agreement have been interpreted as "prudent mate choice" in anticipation of strong assortment by quality (i.e., high-quality individuals prefer high-quality mates, while low-quality individuals prefer low-quality mates; Hardling and Kokko 2005; Burley and Foster 2006; Fawcett and Bleay 2009; Venner et al. 2010). However, our results are incompatible with this interpretation, and suggest instead that the lack of agreement regarding general quality or attractiveness is more profound. Rather than assessing partners in terms of general quality, zebra finches may choose mates based on behavioral compatibility, which appears to be an important determinant of the reproductive success of a pair (Ihle et al. 2015). Pairs that were allowed to form through mutual mate choice achieved a 37% higher fitness than experimentally arranged pairs (Ihle et al. 2015), suggesting that, in this species and under captive conditions, individuals might gain more from mate choice for behavioral compatibility than from directional mate choice for potential indicators of quality. Why would this be? First, quality indicators might only be partly condition-dependent and hence not sufficiently honest indicators of an individual's quality (see the weak correlations in Tables S1 and S2). Second, under social monogamy, selection might not favor high levels of choosiness for mate quality, because the costs related to competition might exceed the benefits of such unidirectional choosiness (Dechaume-Montcharmont et al. 2016). Clearly, in order to understand whether selection favors uniform preferences for high-quality individuals, we would need to quantify both costs and benefits of such choosiness. Both the costs and the benefits of choosiness, as well as mate choice behavior could differ between captive conditions and the wild, but there is no evidence suggesting that zebra finches in the wild pair assortatively for quality indicators.

Broadly speaking, we suggest that the hypothesis that uniform preferences for high-quality individuals have not evolved in some species should not be outright dismissed. Instead, it might be fruitful to explore the possible reasons for a lack of preference for high-quality mates. We emphasize the importance of objectively quantifying the benefits of preferences for quality indicators as well as the strength of such preferences (avoiding publication bias against nonsignificant findings).

DOES PCA ANALYSIS REFLECT VARIATION IN INDIVIDUAL QUALITY?

Nearly all loadings of quality predictors were in the expected direction (Table 2), suggesting that the PCA was successful at summarizing between-individual variation in quality. However, for every single predictor these loadings should be interpreted with caution, because the probability of a loading being in the expected direction by chance lies at $P = 0.5$, and only few predictors were significantly correlated (18 out of 60 relevant pair-wise correlations; Tables S1, S2). Natal brood size was the only predictor that was not associated with quality as expected, neither in males, nor in females. However, a positive rather than a negative relationship between natal brood size and individual quality has also been reported (Kriengwatana et al. 2016). Hence, it is possible that a larger natal brood size is associated with high-quality parents, compensating for the costs of increased sibling competition. Most quality indicators showed only weak loadings on the first principal component (Table 2), suggesting that their value as an indicator is limited. This is also reflected in relatively low Eigenvalues of the first principal component in males and females (around 1.7), even lower than those reported in previous studies using PCA to summarize measures reflecting quality (2.4–2.7; Hamel et al. 2009; Moyes et al. 2009).

QUALITY SCORES REFLECT PAIRING SUCCESS

Our measure of pairing success (sum of the number of days paired to any individual) integrated four aspects: (1) the probability of pairing versus remaining unpaired, (2) the speed of pairing, (3) the ability to maintain the partnership, and (4) the ability to obtain and maintain multiple pair bonds simultaneously. This composite measure of pairing success was robustly related to the PCA scores reflecting variation in individual quality.

There are three possible explanations why PCA scores predicted individual pairing success. (1) Opposite-sex birds may have discriminated against individuals with low PCA scores, leaving them unpaired, and may have preferred individuals with high PCA scores, even if already paired (leading to polygamy). (2) Birds with high PCA scores may be more competitive in interactions with same-sex individuals and hence more likely to be successful in securing one or multiple partners. (3) PCA scores may reflect individual vigor and hence ability and readiness to invest in reproduction, such that the birds in worst condition showed little interest in pairing, whereas the most vigorous birds had enough energy to maintain even multiple pair bonds or care for multiple broods. These nonexclusive alternative explanations are not easy to distinguish, but the first explanation based on discrimination during mate choice appears unlikely in light of the lack of evidence for assortative mating by PCA scores (see below).

The effect of PCA scores on pairing success was stronger in males than in females, which also has multiple possible

explanations. (1) PCA scores in females were based on fewer traits than those in males (Table 2), potentially resulting in a less powerful predictor of pairing success. (2) Intrasexual competition may be more intense in males than in females, leading to low-quality males remaining unpaired, and allowing the highest quality males to become polygynous. Indeed, polygyny was more common than polyandry. (3) Mate choice based on quality indicators might be stronger in females than in males. We have not rigorously assessed whether females prefer males with high PCA scores, but previous work showed no evidence that female zebra finches preferred males with redder beaks (Forstmeier 2004; Forstmeier and Birkhead 2004), or with higher courtship rates (Forstmeier 2004, 2007).

NO EVIDENCE FOR ASSORTATIVE MATING FOR MEASURES OF QUALITY (MUTUAL MATE CHOICE)

We found no evidence for assortative pairing by PC scores or by the underlying traits despite high statistical power. Because power is a matter of the expected effect size, which could take any value, a comparison to the human mate choice literature is insightful. In humans, body height is one of the less important criteria for mate choice, but directional preferences and mutual mate choice still result in a Pearson correlation coefficient of $r = 0.23$ (meta-analysis of 154 estimates; Stulp et al. 2017). In our study, the power to detect a correlation of this magnitude lies at 0.99, even when focusing only on the 292 monogamous pair bonds that lasted for the entire breeding round. When including all 423 pairs, our study reached a power of 95% to detect correlations above $r = 0.16$.

For such positive correlation to arise from mutual mate choice, there would have to be a reasonably high degree of between-individual agreement about what constitutes an attractive partner. Yet, both male and female zebra finches show only low levels of such agreement (Forstmeier and Birkhead 2004; Wang et al. 2017), meaning that their preferences are highly individualistic or flexible. Such individual mating preferences imply that intrasexual competition for the presumed highest quality mates is reduced and that most individuals may be able to pair with their preferred mate and achieve maximal fitness through effective cooperation in a lifelong pair bond (Ihle et al. 2015). In contrast, under the conventional scenario with consensus in mate preferences, all members of a sex compete for the same (few) high-quality partners, that is for those potential mates that show the highest values for quality indicator traits. Under such conditions, most if not all individuals would pay a cost for the intense competition, while the successful competitors only achieve a relatively small benefit from being choosy, unless the quality indicators strongly and reliably predict fitness (but see the relatively weak loadings in Table 2). Moreover, unsuccessful competitors would end up unpaired or paired to nonpreferred mates, which may result in unstable partnerships that suffer (in terms of fitness) from a lack of mutual commitment (Ihle et al. 2015).

The above discussion of zebra finch mate choice patterns as hypothetically adaptive might appear at odds with the observation that several individuals in our experiments remained unpaired (Fig. 1), and hence ended up with zero or low reproductive success. For some individuals this might simply be explained by poor condition, for instance due to inbreeding depression. Alternatively, some individuals might have skipped the opportunity to reproduce as a consequence of being too choosy (i.e., unwilling to pair with the left-over individual(s) of the opposite sex). This behavior seems maladaptive in the captive context, but a limited availability of potential mates may be rare in the wild in this species.

NO EVIDENCE FOR DIFFERENTIAL ALLOCATION BASED ON QUALITY DIFFERENCES

We found no evidence for the hypothesis that low-quality individuals would increase their own value as a partner by taking a greater share in parental care, thereby securing their higher quality partner. This means that there was no positive differential allocation, even though pairs were often greatly mismatched for quality. However, in this study we did not allow eggs to hatch (all eggs were replaced by dummy eggs and were removed after 10 days of incubation), and hence the most demanding part of brood care (feeding the offspring) was not measured. Nevertheless, if differential allocation by low-quality individuals is a signal that is effective in retaining a higher quality partner, one would expect that an increased readiness to invest in care should be signaled early during the first reproductive event, when pair bonds are still fragile (divorce becomes rare after being paired for about three weeks). Given that incubation at an ambient temperature of about 20°C is energetically not very demanding (Vleck 1981), such signaling should have been possible even for individuals in relatively poor condition. Yet, low-quality individuals within mismatched pairs did not seem to make an effort to signal their parental qualities to secure their high-quality partner. Given the absence of assortative mating for quality, the most parsimonious explanation for the lack of differential allocation might be that the risk of losing a partner is equal for matched and mismatched pairs, because individual quality is not a target of zebra finch mate choice.

Our observation that high-quality males were more likely to become polygynous is also worth discussing in this context. Burley (1988) also observed it and argued that due to positive differential allocation by the lower quality partner, high-quality individuals would have to do less of the parental care in the first brood (with the primary partner investing more), allowing them to invest more in attracting a second partner. Because we found no evidence for positive differential allocation (reduced care by high-quality males) within monogamous pairs, this explanation seems unlikely. Instead, we suggest that high-quality males simply seek

additional mates because they have more energy to spend than low-quality males.

THE IMPORTANCE OF UNBIASED QUANTIFICATION OF SIGNAL HONESTY AND PREFERENCE STRENGTH

Research on mate choice has often been carried out with the aim to identify male traits under strong directional selection (Andersson and Simmons 2006). Typically, several potential male traits have been studied and preferences or choice outcomes have been measured in numerous ways. Often, this may have resulted in a considerable multiple testing problem and probably led to selective reporting of positive findings rather than comprehensive reporting of all tests with loss of significance after Bonferroni adjustment (Forstmeier et al. 2016). The strive for statistical significance has created a mate choice literature that is heavily biased toward inflated effect size estimates and that shows considerable heterogeneity in estimates (i.e., failure to replicate) (Parker 2013). Hence, to determine whether males and females choose mates based on quality indicators and–-ultimately–-to understand the evolution of mate choice based on individual quality, we need to shift our efforts toward unbiased quantification of some key parameters. This would require complete and unbiased reporting of all relevant parameters that have been examined, or at least reporting an unbiased estimate of the average effect size within a study (e.g., across multiple ornaments or across multiple ways of analysis; Forstmeier et al. 2016). Such unbiased estimates should be obtained for (1) the degree of honesty of signals (i.e., how well expression of the signal reflects individual quality), (2) the strength of preferences for these signals, and (3) the costs of competition for mates. This would allow us to examine whether there is currently selection for strong directional preferences, that is whether the benefits of choosiness exceed the costs.

AUTHOR CONTRIBUTIONS

WF, BK and DW contributed to the design of the research. DW and WF measured ornaments and recorded songs. DW contributed to nest checks and observations of pairing and parental care. WF and DW analyzed the data. DW, WF and BK wrote the paper.

ACKNOWLEDGMENTS

We thank Katrin Martin for extensive help with the breeding experiments, Melanie Schneider for molecular work, and Sonja Bauer, Andrea Kortner, Jane Didsbury, and Petra Neubauer for animal care. Pietro d'Amelio kindly helped with song recording and analysis. We are grateful to Michael Jennions and an anonymous reviewer for constructive comments on the manuscript.

This work was supported by the Max Planck Society (to B.K.) and the China Scholarship Council (CSC; stipend to D.W.).

DATA ARCHIVING

Data available from the Dryad Digital Repository http://datadryad.org/resource/doi:10.5061/dryad.85950 (Wang et al. 2017)

LITERATURE CITED

Abramoff, M. D., P. J. Magalhães, and S. J. Ram. 2004. Image processing with ImageJ. Biophotonics int. 11:36–42.

Adkins-Regan, E. 1998. Hormonal mechanisms of mate choice. Am. Zool. 38:166–178.

———. 2007. Hormones and the development of sex differences in behavior. J. Ornithol. 148:S17–S26.

———. 2011. Neuroendocrine contributions to sexual partner preference in birds. Front Neuroendocrinol. 32:155–163.

Andersson, M. 1982. Female choice selects for extreme tail length in a widowbird. Nature 299:818–820.

Andersson, M., and L. W. Simmons. 2006. Sexual selection and mate choice. Trends Ecol. Evol. 21:296–302.

Andersson, M. B. 1994. Sexual selection. Princeton Univ. Press, Princeton, N.J.

Baldauf, S. A., H. Kullmann, S. H. Schroth, T. Thunken, and T. C. M. Bakker. 2009. You can't always get what you want: size assortative mating by mutual mate choice as a resolution of sexual conflict. Bmc Evol. Biol. 9:129.

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

Bergeron, P., R. Baeta, F. Pelletier, D. Reale, and D. Garant. 2011. Individual quality: tautology or biological reality? J. Anim. Ecol. 80:361–364.

Bergstrom, C. T., and L. A. Real. 2000. Towards a theory of mutual mate choice: lessons from two-sided matching. Evol. Ecol. Res. 2:493–508.

Birkhead, T. R., E. J. Pellatt, I. M. Matthews, N. J. Roddis, F. M. Hunter, F. McPhie, and H. Castillo-Juarez. 2006. Genic capture and the genetic basis of sexually selected traits in the zebra finch. Evolution 60:2389–2398.

Blount, J. D., N. B. Metcalfe, T. R. Birkhead, and P. F. Surai. 2003. Carotenoid modulation of immune function and sexual attractiveness in zebra finches. Science 300:125–127.

Bolund, E., K. Martin, B. Kempenaers, and W. Forstmeier. 2010a. Inbreeding depression of sexually selected traits and attractiveness in the zebra finch. Anim. Behav. 79:947–955.

Bolund, E., H. Schielzeth, and W. Forstmeier. 2010b. No heightened condition dependence of zebra finch ornaments—a quantitative genetic approach. J. Evol. Biol. 23:586–597.

Bolund, E., H. Schielzeth, and W. Forstmeier. 2011. Correlates of male fitness in captive zebra finches—a comparison of methods to disentangle genetic and environmental effects. Bmc Evol. Biol. 11:327.

Boogert, N. J., L. A. Giraldeau, and L. Lefebvre. 2008. Song complexity correlates with learning ability in zebra finch males. Anim. Behav. 76:1735–1741.

Burley, N. 1983. The meaning of assortative mating. Ethol. Sociobiol. 4:191–203.

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

Burley, N. T., and V. S. Foster. 2006. Variation in female choice of mates: condition influences selectivity. Anim. Behav. 72:713–719.

Catchpole, C. K., and P. J. B. Slater. 2008. Bird song: Biological themes and variations. Cambridge university press. 2nd ed. pp. 1–335.

Chapman, J. R., S. Nakagawa, D. W. Coltman, J. Slate, and B. C. Sheldon. 2009. A quantitative review of heterozygosity-fitness correlations in animal populations. Mol. Ecol. 18:2746–2765.

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

Cotton, S., K. Fowler, and A. Pomiankowski. 2004. Do sexual ornaments demonstrate heightened condition-dependent expression as predicted by the handicap hypothesis? P Roy Soc. B Biol. Sci. 271:771–783.

Courtiol, A., L. Etienne, R. Feron, B. Godelle, and F. Rousset. 2016. The evolution of mutual mate choice under direct benefits. Am. Nat. 188:521–538.

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.

DeKogel, C. H., and H. J. Prijs. 1996. Effects of brood size manipulations on sexual attractiveness of offspring in the zebra finch. Anim. Behav. 51:699–708.

Doutrelant, C., A. Gregoire, A. Midamegbe, M. Lambrechts, and P. Perret. 2012. Female plumage coloration is sensitive to the cost of reproduction. An experiment in blue tits. J. Anim. Ecol. 81:87–96.

Dunn, P. O., J. C. Garvin, L. A. Whittingham, C. R. Freeman-Gallant, and D. Hasselquist. 2010. Carotenoid and melanin-based ornaments signal similar aspects of male quality in two populations of the common yellowthroat. Funct. Ecol. 24:149–158.

Fawcett, T. W., and C. Bleay. 2009. Previous experiences shape adaptive mate preferences. Behav. Ecol. 20:68–78.

Fawcett, T. W., and R. A. Johnstone. 2003. Mate choice in the face of costly competition. Behav. Ecol. 14:771–779.

Forstmeier, W. 2004. Female resistance to male seduction in zebra finches. Anim. Behav. 68:1005–1015.

———. 2007. Do individual females differ intrinsically in their propensity to engage in extra-pair copulations? PLoS One 2:e952.

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., C. Burger, K. Temnow, and S. Deregnaucourt. 2009. The genetic basis of zebra finch vocalizations. Evolution 63:2114–2130.

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., H. Schielzeth, J. C. Mueller, H. Ellegren, and B. Kempenaers. 2012. Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. Mol. Ecol. 21:3237–3249.

Forstmeier, W., G. Segelbacher, J. C. Mueller, and B. Kempenaers. 2007. Genetic variation and differentiation in captive and wild zebra finches (*Taeniopygia guttata*). Mol. Ecol. 16:4039–4050.

Forstmeier, W., E.-J. Wagenmakers, and T. H. Parker. 2016. Detecting and avoiding likely false-positive findings—a practical guide. Biol Rev. https://doi.org/10.1111/brv.12315

Galipaud, M., L. Bollache, and F. X. Dechaume-Moncharmont. 2013. Assortative mating by size without a size-based preference: the female-sooner norm as a mate-guarding criterion. Anim Behav 85:35–41.

Gimelfarb, A. 1988a. Processes of Pair Formation Leading to Assortative Mating in Biological Populations - Dynamic Interaction-Model. Theor Popul Biol 34:1–23.

Gimelfarb, A. 1988b. Processes of Pair Formation Leading to Assortative Mating in Biological Populations - Encounter-Mating Model. Am Nat 131:865–884.

Griffith, S. C., and K. L. Buchanan. 2010. The Zebra Finch: the ultimate Australian supermodel. Emu 110:V–xii.

Hamel, S., J. M. Gaillard, M. Festa-Bianchet, and S. D. Cote. 2009. Individual quality, early-life conditions, and reproductive success in contrasted populations of large herbivores. Ecology 90:1981–1995.

Hamilton, W. D., and M. Zuk. 1982. Heritable true fitness and bright birds—a role for parasites. Science 218:384–387.

Hardling, R., and H. Kokko. 2005. The evolution of prudent choice. Evol. Ecol. Res. 7:697–715.

Hauber, M. E., D. L. M. Campbell, and S. M. N. Woolley. 2010. The functional role and female perception of male song in zebra finches. Emu 110:209–218.

Holveck, M. J., and K. Riebel. 2010. Low-quality females prefer low-quality males when choosing a mate. P Roy Soc. B Biol. Sci. 277:153–160.

Hooper, P. L., and G. F. Miller. 2008. Mutual mate choice can drive costly signaling even under perfect monogamy. Adapt. Behav. 16:53–70.

Houtman, A. M. 1992. Female zebra finches choose extra-pair copulations with genetically attractive males. P Roy Soc. B Biol. Sci. 249:3–6.

Ihle, M., B. Kempenaers, and W. Forstmeier. 2013. Does hatching failure breed infidelity? Behav. Ecol. 24:119–127.

———. 2015. Fitness benefits of mate choice for compatibility in a socially monogamous species. PLoS Biol 13:e1002248. https://doi.org/10.1371/journal.pbio.1002248

Jiang, Y. X., D. I. Bolnick, and M. Kirkpatrick. 2013. Assortative mating in animals. Am. Nat. 181:E125–E138.

Johnstone, R. A. 1997. The tactics of mutual mate choice and competitive search. Behav. Ecol. Sociobiol. 40:51–59.

Johnstone, R. A., J. D. Reynolds, and J. C. Deutsch. 1996. Mutual mate choice and sex differences in choosiness. Evolution 50:1382–1391.

Jones, I. L., and F. M. Hunter. 1993. Mutual sexual selection in a monogamous seabird. Nature 362:238–239.

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

Kokko, H., R. Brooks, M. D. Jennions, and J. Morley. 2003. The evolution of mate choice and mating biases. P Roy Soc B Biol. Sci. 270:653–664.

Kokko, H., and M. D. Jennions. 2008. Parental investment, sexual selection and sex ratios. J. Evol. Biol. 21:919–948.

Kokko, H., and R. A. Johnstone. 2002. Why is mutual mate choice not the norm? Operational sex ratios, sex roles and the evolution of sexually dimorphic and monomorphic signalling. Philos. Trans. R Soc. B Biol. Sci. 357:319.

Kriengwatana, B., M. J. Spierings, and C. ten Cate. 2016. Auditory discrimination learning in zebra finches: effects of sex, early life conditions and stimulus characteristics. Anim. Behav. 116:99–112.

Kuijper, B., I. Pen, and F. J. Weissing. 2012. A guide to sexual selection theory. Ann. Rev. Ecol. Evol. Syst. 43:287–+.

Lailvaux, S. P., and M. M. Kasumovic. 2011. Defining individual quality over lifetimes and selective contexts. P Roy Soc. B Biol. Sci. 278:321–328.

Lindstrom, J. 1999. Early development and fitness in birds and mammals. Trends Ecol. Evol. 14:343–348.

Mathot, K. J., K. Martin, B. Kempenaers, and W. Forstmeier. 2013. Basal metabolic rate can evolve independently of morphological and behavioural traits. Heredity 111:175–181.

Moyes, K., B. J. T. Morgan, A. Morris, S. J. Morris, T. H. Clutton-Brock, and T. Coulson. 2009. Exploring individual quality in a wild population of red deer. J. Anim. Ecol. 78:406–413.

Naguib, M., and A. Nemitz. 2007. Living with the past: nutritional stress in juvenile males has immediate effects on their plumage ornaments and on adult attractiveness in zebra finches. PLoS ONE 2:e901. https://doi.org/10.1371/journal.pone.0000901

Nakagawa, S., N. Ockendon, D. O. S. Gillespie, B. J. Hatchwell, and T. Burke. 2007. Assessing the function of house sparrows' bib size using a flexible meta-analysis method. Behav. Ecol. 18:831–840.

Nilsson, J. A., and E. Svensson. 1996. The cost of reproduction: a new link between current reproductive effort and future reproductive success. Proc. Roy Soc. B Biol. Sci. 263:711–714.

Parker, T. H. 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.

Pincemy, G., F. S. Dobson, and P. Jouventin. 2009. Experiments on colour ornaments and mate choice in king penguins. Anim. Behav. 78:1247–1253.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

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

Reid, J. M., P. Arcese, A. L. E. V. Cassidy, S. M. Hiebert, J. N. M. Smith, P. K. Stoddard, A. B. Marr, and L. F. Keller. 2004. Song repertoire size predicts initial mating success in male song sparrows, *Melospiza melodia*. Anim. Behav. 68:1055–1063.

Riebel, K. 2003. Developmental influences on auditory perception in female zebra finches—is there a sensitive phase for song preference learning? Anim. Biol. 53:73–87.

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

Ritschard, M., K. Riebel, and H. Brumm. 2010. Female zebra finches prefer high-amplitude song. Anim. Behav. 79:877–883.

Roberts, M. L., K. L. Buchanan, A. T. D. Bennett, and M. R. Evans. 2007. Mate choice in zebra finches: does corticosterone play a role? Anim. Behav. 74:921–929.

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

Soma, M., and L. Z. Garamszegi. 2011. Rethinking birdsong evolution: meta-analysis of the relationship between song complexity and reproductive success. Behav. Ecol. 22:363–371.

Spencer, K. A., K. L. Buchanan, A. R. Goldsmith, and C. K. Catchpole. 2003. Song as an honest signal of developmental stress in the zebra finch (*Taeniopygia guttata*). Horm. Behav. 44:132–139.

Spencer, K. A., J. H. Wimpenny, K. L. Buchanan, P. G. Lovell, A. R. Goldsmith, and C. K. Catchpole. 2005. Developmental stress affects the attractiveness of male song and female choice in the zebra finch (*Taeniopygia guttata*). Behav. Ecol. Sociobiol. 58:423–428.

Stulp, G., M. J. P. Simons, S. Grasman, and T. V. Pollet. 2017. Assortative mating for human height: a meta-analysis. American Journal of Human Biology. 29:e22917.

Tchernichovski, O., T. J. Lints, S. Deregnaucourt, A. Cimenser, and P. P. Mitra. 2004. Studying the song development process rationale and methods. Ann. NY Acad. Sci. 1016:348–363.

ten Cate, C., and D. R. Vos. 1999. Sexual imprinting and evolutionary processes in birds: a reassessment. Adv. Study Behav. 28:1–31.

Tschirren, B., E. Postma, A. N. Rutstein, and S. C. Griffith. 2012. When mothers make sons sexy: maternal effects contribute to the increased sexual attractiveness of extra-pair offspring. Proc. Roy Soc. B Biol. Sci. 279:1233–1240.

Tschirren, B., A. N. Rutstein, E. Postma, M. Mariette, and S. C. Griffith. 2009. Short- and long-term consequences of early developmental conditions: a case study on wild and domesticated zebra finches. J. Evol. Biol. 22:387–395.

Van Noordwijk, A. J., and G. de Jong. 1986. Acquisition and allocation of resources: their influence on variation in life history tactics. Am. Nat. 128:137–142.

Venner, S., C. Bernstein, S. Dray, and M.-C. Bel-Venner. 2010. Make love not war: when should less competitive males choose low-quality but defendable females? Am. Nat. 175:650–661.

Vleck, C. M. 1981. Energetic cost of incubation in the zebra finch. Condor 83:229–237.

Vyas, A., C. Harding, J. McGowan, R. Snare, and D. Bogdan. 2008. Noradrenergic neurotoxin, N-(2-chloroethyl)-N-ethyl-2-bromobenzylamine hydrochloride (DSP-4), treatment eliminates estrogenic effects on,song responsiveness in female zebra finches (*Taeniopygia guttata*). Behav. Neurosci. 122:1148–1157.

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

Welch, A. M., R. D. Semlitsch, and H. C. Gerhardt. 1998. Call duration as an indicator of genetic quality in male gray tree frogs. Science 280:1928–1930.

Wells, S. J., W. H. Ji, J. Dale, B. Jones, and D. Gleeson. 2015. Male size predicts extrapair paternity in a socially monogamous bird with extreme sexual size dimorphism. Behav. Ecol. 26:200–206.

Wilson, A. J., and D. H. Nussey. 2010. What is individual quality? An evolutionary perspective. Trends Ecol. Evol. 25:207–214.

Woodgate, J. L., S. Leitner, C. K. Catchpole, M. L. Berg, A. T. D. Bennett, and K. L. Buchanan. 2011. Developmental stressors that impair song learning in males do not appear to affect female preferences for song complexity in the zebra finch. Behav. Ecol. 22:566–573.

Woodgate, J. L., M. M. Mariette, A. T. D. Bennett, S. C. Griffith, and K. L. Buchanan. 2012. Male song structure predicts reproductive success in a wild zebra finch population. Anim. Behav. 83:773–781.

Zann, R. A. 1996. The zebra finch: a synthesis of field and laboratory studies. Oxford Univ. Press, Oxford; New York.

Associate Editor: E. Derryberry
Handling Editor: P. Tiffin

Supporting Information

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

Fig. S1. Average PC scores (± SE) of monogamous, polygamous and unpaired individuals of both sexes (black bars are females, grey bars are males).
Fig. S2. Principal component scores of mated pairs of different pairing status (as in Fig. 2, but both breeding rounds combined).
Fig. S3. Male and female quality (PC scores from the other breeding round) and level of assortative mating for PC scores (Pearson correlation coefficient r) in relation to the order of pairing within aviaries.
Table S1. Pairwise Pearson's correlation coefficients among traits related to female quality.
Table S2. Pairwise Pearson's correlation coefficients among traits related to male quality.

No mutual mate choice for quality in zebra finches: time to question a widely-held assumption

Daiping Wang, Wolfgang Forstmeier, Bart Kempenaers,

SUPPLEMENT

Fig. S1. Average PC scores (± SE) of monogamous, polygamous and unpaired individuals of both sexes (black bars are females, grey bars are males). Sample sizes (number of individuals) are indicated. (a) Pairing status during the first breeding round in relation to the PC score from the second breeding round (b) Pairing status from the second round in relation to the PC score from the first round. PC scores of polyandrous females were significantly higher than those of unpaired females in (b) $t = 2.52$, $p = 0.01$, but not in (a) $t = 1.63$, $p = 0.11$. PC scores of polygynous males were significantly higher than those of monogamous males (a) $t = 5.71$, $p < 0.001$, (b) $t = 5.01$, $p < 0.001$.

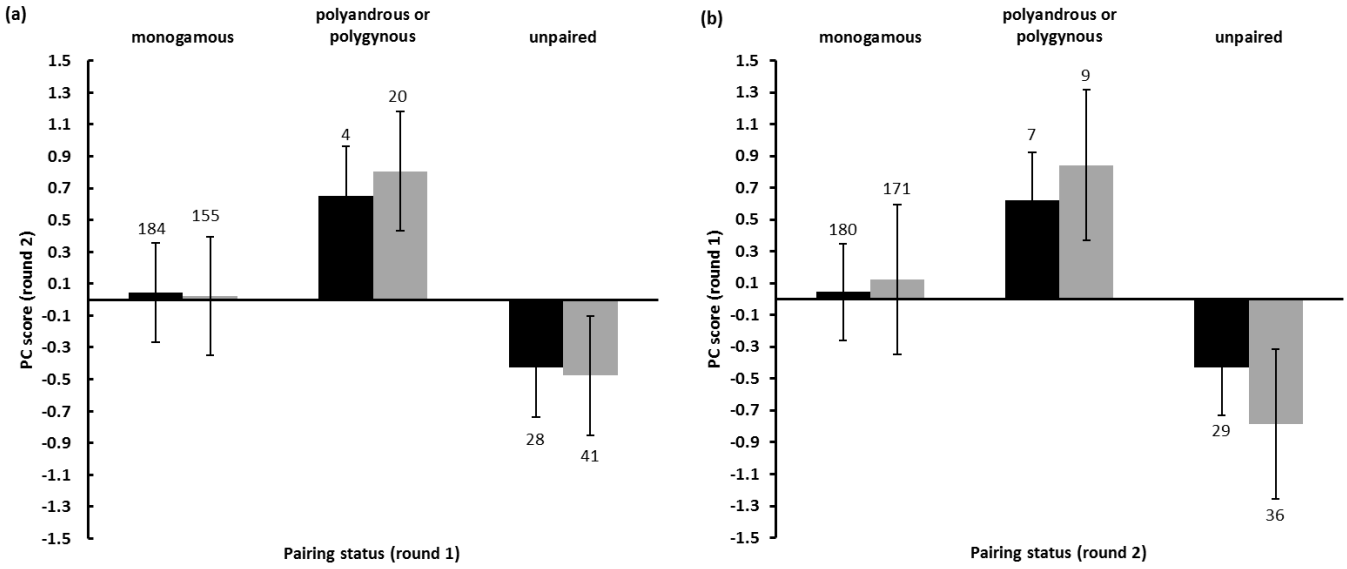


Fig. S2. Principal component scores of mated pairs of different pairing status (as in Fig. 2, but both breeding rounds combined). Blue diamonds show monogamous pairs that lasted to the end of a breeding round (n = 292). (a) Partners of polygynous males are shown as red squares (primary females, n = 29) or green triangles (secondary females, n = 33). (b) Partners of polyandrous females are shown as red squares (primary males, n = 11) or green triangles (secondary males , n = 11).

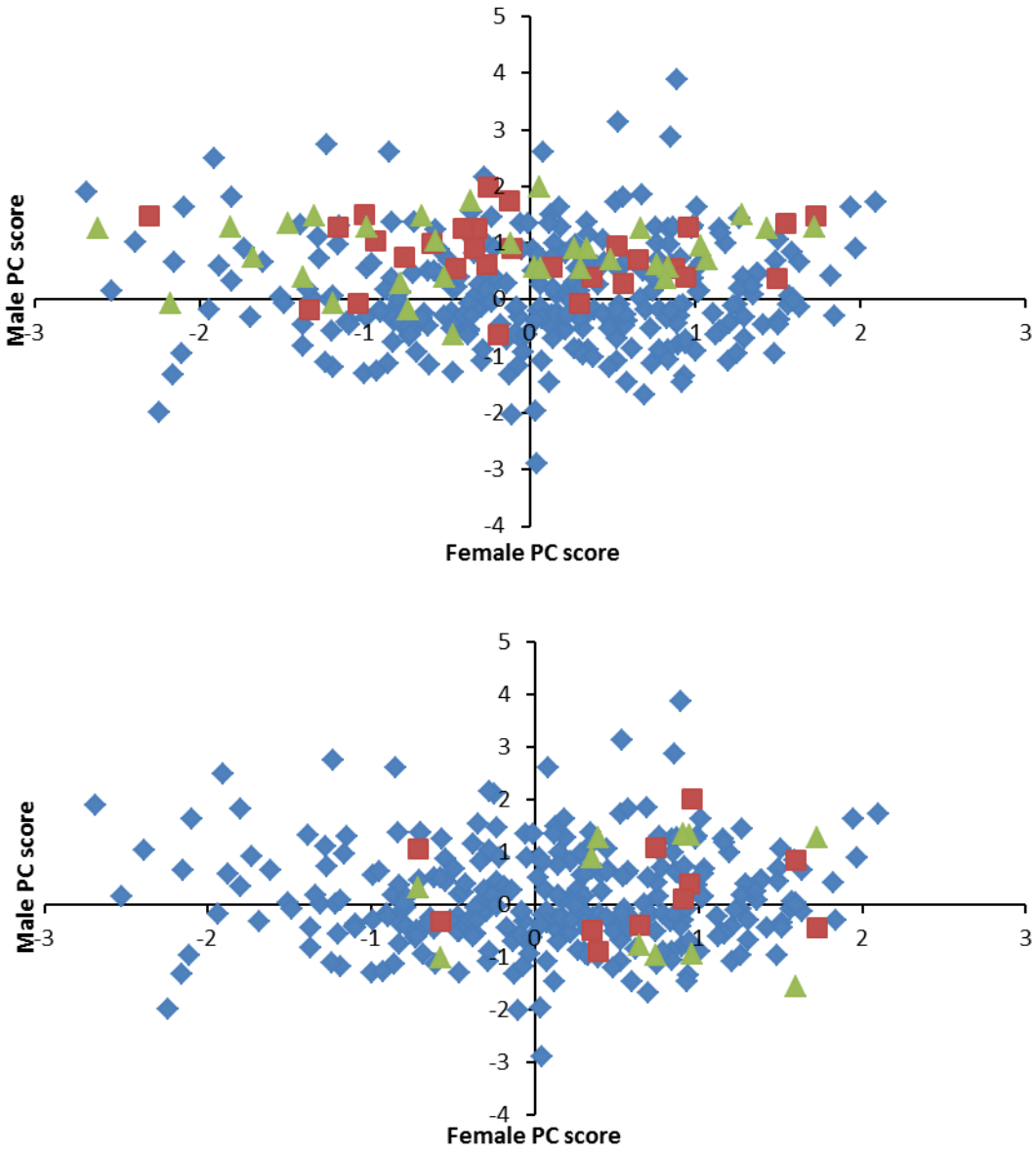


Fig. S3. Male and female quality (PC scores from the other breeding round) and level of assortative mating for PC scores (Pearson correlation coefficient r) in relation to the order of pairing within aviaries. All pair bonds that were formed within an aviary were ranked by the order of pair formation (Pairing order within aviary). Due to changes in pair bonds (divorce and subsequent re-pairing with another individual) there were up to eight pair bonds recorded within an aviary (rather than maximally six, given 6 males and 6 females per aviary). Sample sizes (number of pair bonds within each rank) are indicated. In cases of ties (same time of pairing for two pairs) both pairs were given the same lower rank, explaining why there are n = 82 first pair bonds with only 72 aviaries. Mean values ± 95% CI are indicated.

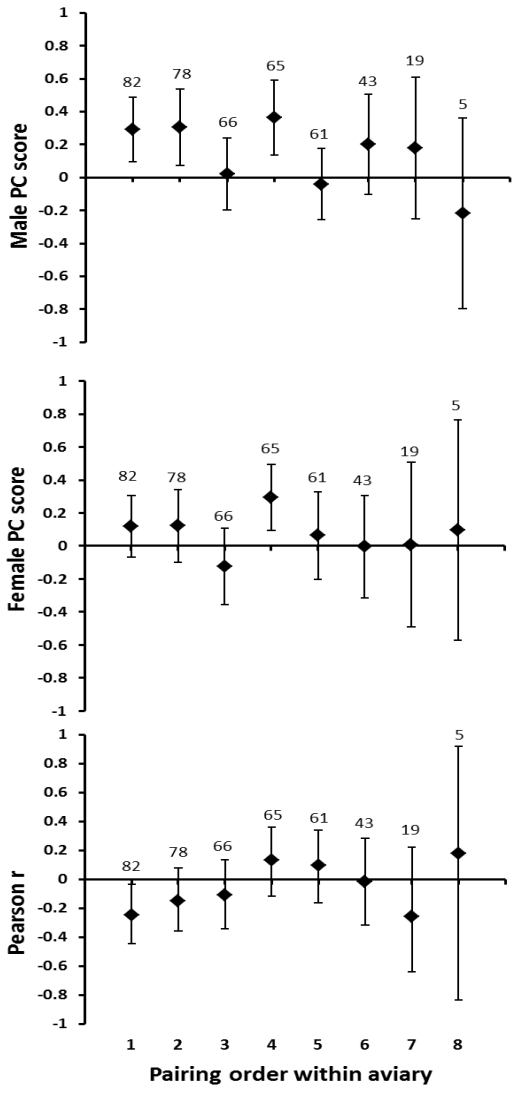


Table S1. Pairwise Pearson's correlation coefficients among traits related to female quality. Sample size (N) is indicated and asteriscs mark significant relationships (without correction for multiple testing) *: $p < 0.05$, **: $p < 0.01$, two-tailed.

	Fecundity round 2	Mean fecundity	Inbreeding coefficient F	Mass day 8	Mass day 100	Natal brood size	Beak color
Fecundity round 1	0.632**	0.898**	-0.076	0.206**	0.161*	0.075	-0.002
N	213	216	216	216	216	216	216
Fecundity round 2		0.913**	-.144*	0.164*	0.201**	0.028	0.024
N		216	216	216	216	216	216
Mean fecundity			-0.110	0.199**	0.198**	0.055	0.012
N			219	219	219	219	219
Inbreeding coefficient F				-0.087	-0.128	0.005	0.154*
N				219	219	219	219
Mass day 8					0.432**	-0.049	0.228**
N					219	219	219
Mass day 100						-0.067	0.210**
N						219	219
Natal brood size							0.144*
N							219

Table S2. Pairwise Pearson's correlation coefficients among traits related to male quality. Sample size (N) is indicated and asteriscs mark significant relationships (without correction for multiple testing) *: $p < 0.05$, **: $p < 0.01$, two-tailed.

	Siring success round 2	Mean siring success	Inbreeding coefficient F	Mass day 8	Mass day 100	Natal brood size	Cheek patch size	Beak color	Amplitude	Repertoire size	Courtship rate
Siring success round 1	0.457**	0.842**	-0.228**	0.081	0.068	0.143*	0.066	-0.097	0.244**	0.072	0.075
N	215	216	216	216	216	216	200	216	190	190	216
Siring success round 2		0.865**	-.236**	0.075	0.104	0.104	-0.005	-0.046	0.070	0.107	0.064
N		215	215	215	215	215	200	215	190	190	215
Mean siring success			-.267**	0.095	0.105	0.142*	0.034	-0.088	0.183*	0.107	0.078
N			216	216	216	216	200	216	190	190	216
Inbreeding coefficient F				0.072	-0.042	-0.017	-0.004	0.044	-0.067	-0.021	-0.067
N				216	216	216	200	216	190	190	216
Mass day 8					0.233**	-0.130	0.172*	0.061	0.014	0.098	-0.033
N					216	216	200	216	190	190	216
Mass day 100						0.057	0.156*	0.128	0.077	0.046	0.037
N						216	200	216	190	190	216
Natal brood size							0.219**	0.118	0.236**	0.127	0.003
N							200	216	190	190	216
Cheek patch size								0.093	0.165*	-0.081	0.111
N								200	190	190	200
Beak color									-0.149*	-0.023	0.183**
N									190	190	216
Amplitude										-0.094	0.000
N										190	190
Repertoire size											-0.045
N											190

Chapter 3: Irreproducible text-book 'knowledge': the effects of color bands on zebra finch fitness

Short title: Color bands have no effect on fitness in zebra finches

Abstract: 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 8 out 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.

Published as: Daiping Wang, Wolfgang Forstmeier, Malika Ihle, Mehdi Khadraoui, Sofia Jerónimo, Katrin Martin and Bart Kempenaers 2018: Irreproducible text‐book "knowledge": The effects of color bands on zebra finch fitness. Evolution 72(4):961-976.

VOLUME 72, NUMBER 4, APRIL 2018

EVOLUTION

INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

SSE
SOCIETY for the STUDY of EVOLUTION

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

© 2018 The Author(s). *Evolution* © 2018 The Society for the Study of Evolution.

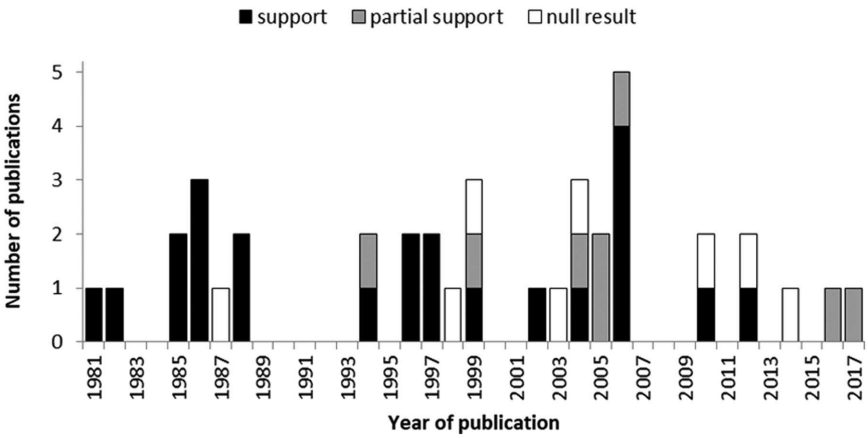


Figure 1. **Summary of publications ($n = 39$) of experiments in which male zebra finches were fitted with red versus green color bands. Shown are the number of studies and their year of publication. Studies were classified as (1) providing support ($n = 23$) for the hypothesis that red-banded males 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 non-independence. 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

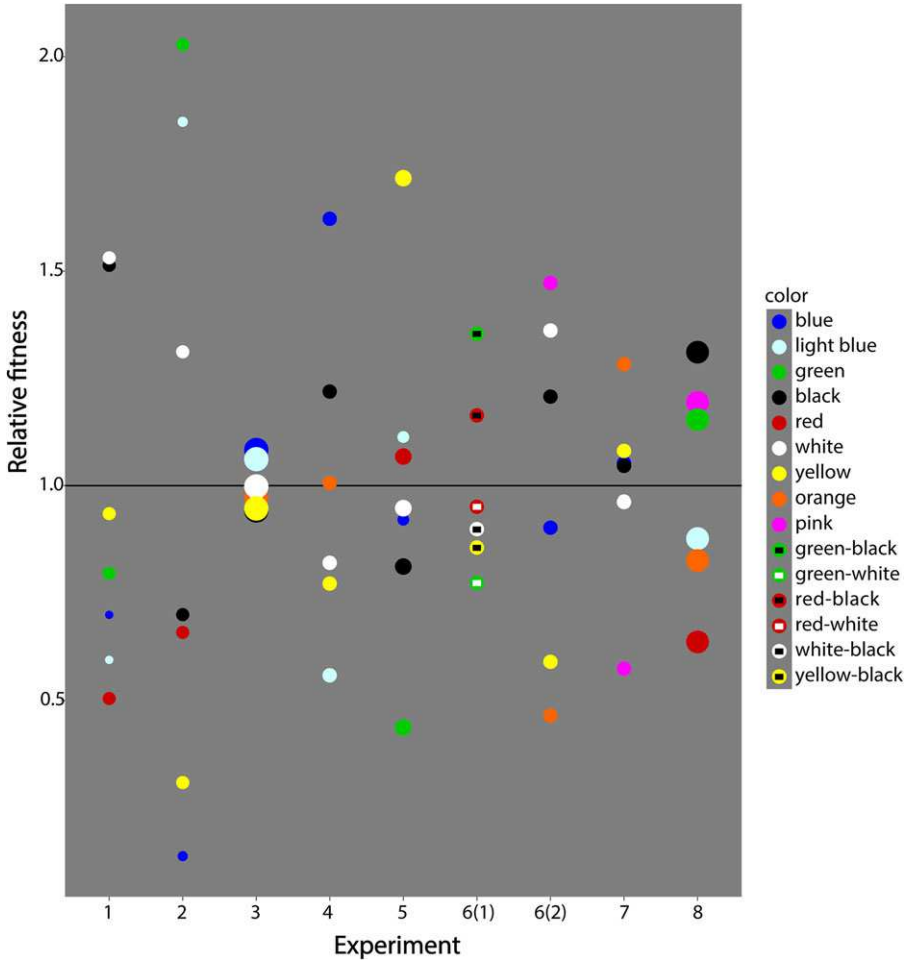


Figure 2. **Mean relative fitness of female zebra finches by band color for each of eight experiments. Each dot represents the average relative fitness (number of independent offspring) of all females with that color band. The size of the dots reflects sample size (number of females ranging from 2 to 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

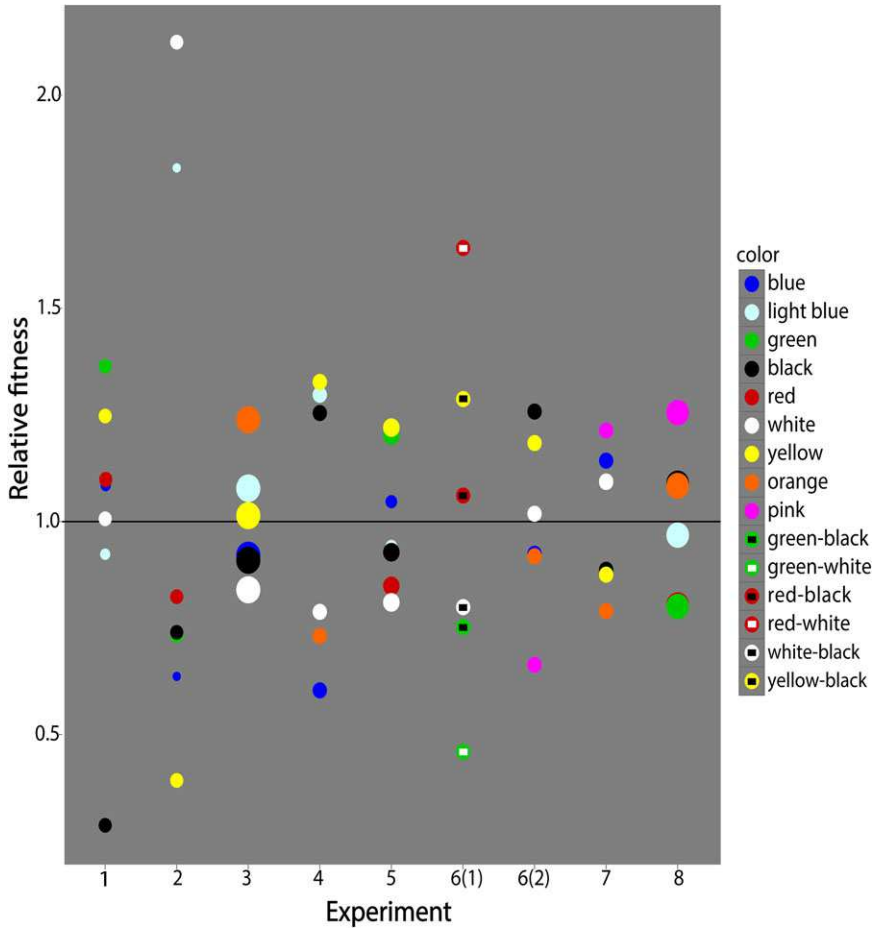


Figure 3. **Mean relative fitness of male zebra finches by band color for each of eight experiments. Each dot represents the average relative fitness (number of independent offspring sired) of all males with that color band. The size of the dots reflects sample size (number of males ranging from 2 to 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

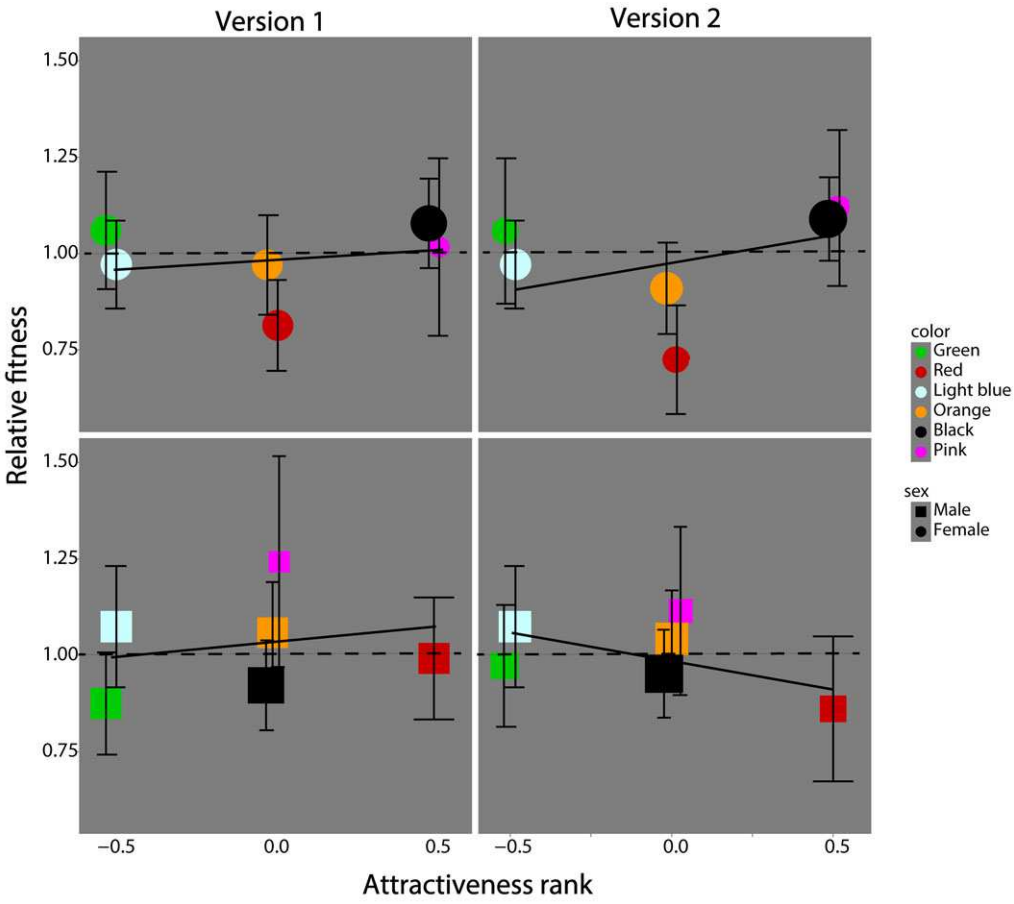


Figure 4. **Regression of mean relative fitness of female (top row) and male zebra finches (bottom row) across all eight experiments as a function of the suggested attractiveness rank of each band color (based on the literature, see Introduction and Methods). Attractive colors were coded as +0.5, unattractive 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.

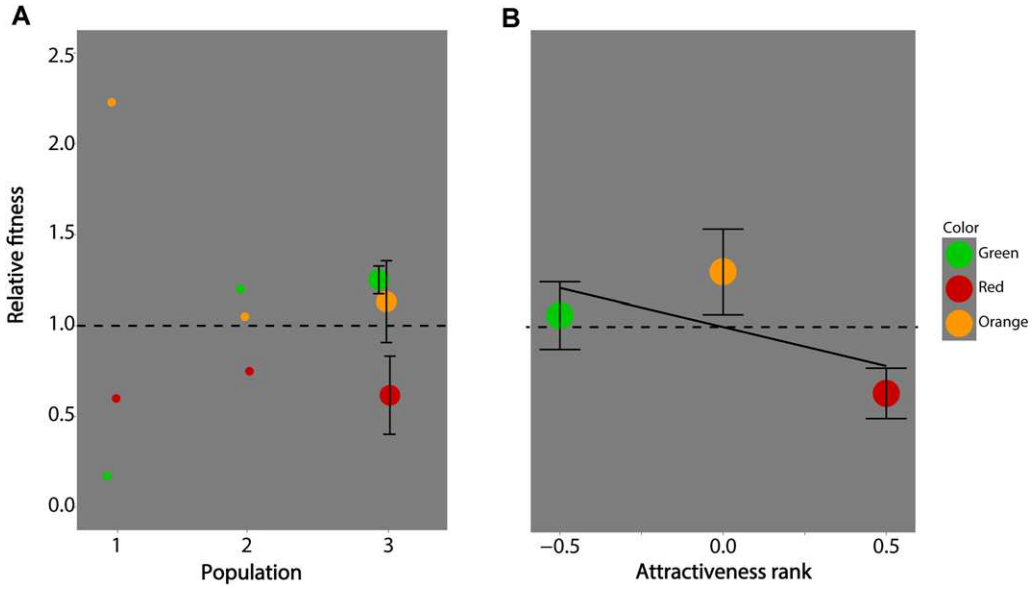


Figure 5. **Summary of results from other laboratories (von Engelhardt 2004). (A) Mean relative fitness of male zebra finches as a function of band color in three captive populations (1: data collected by K. Witte in Bielefeld, (2) R. Zann in Melbourne, (3) N. von Engelhardt in Groningen). Each dot represents 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

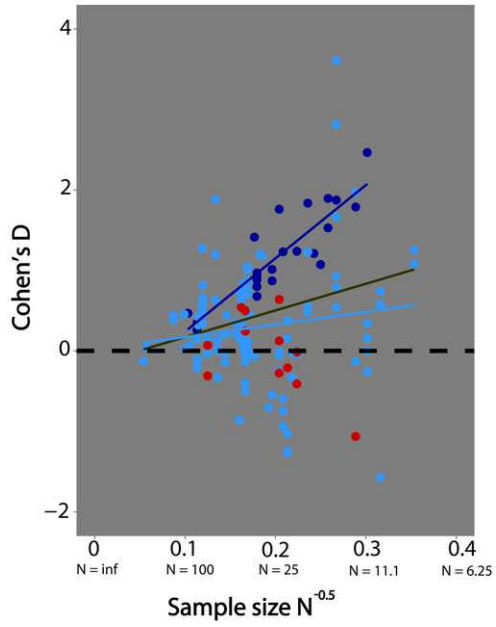


Figure 6. Funnel plot showing published effect size estimates (Cohen's *D* for red vs. green color bands, $n = 141$) in relation to their sample size. The *x*-axis shows sample size $N^{-0.5}$, where *N* is the total number of males (red plus green), or offspring (of red plus green males), or females (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).

Irreproducible text-book 'knowledge': the effects of color bands on zebra finch fitness

Daiping Wang, Wolfgang Forstmeier, Malika Ihle, Mehdi Khadraoui, Sofia Jerónimo, Katrin Martin and Bart Kempenaers

SUPPLEMENT

Additional Methods

Origin of study populations

(1) Population 'Melbourne' originated from birds caught in the wild about three to four generations ago, with 40 males and 40 females exported to Seewiesen, Germany, in December 2015.

(2) Population 'Bielefeld' (described as population#4 in [1]) was derived from wild-caught birds from northern Victoria about 12-15 generations ago. In 1992, 12 males and 12 females had been exported to Bielefeld, Germany, and bred there. In 2009, 109 individuals were transferred from Bielefeld to Seewiesen, where the population has been maintained since.

(3) Population 'Krakow' consists of F1 and F2 hybrids between two domesticated European populations, namely birds from Krakow, Poland (population # 11 in [1]) and birds from Sheffield, UK (see 'Seewiesen' below; population # 18 in [1]). A total of 25 males and 25 females were transferred from Krakow to Seewiesen, in 2011 and in 2013.

(4) Population 'Seewiesen' originates from about 450 birds that were brought from Sheffield, UK, in 2004. This domesticated population has been maintained at Seewiesen since then

(generations F1 to F4). In 2009, we initiated the breeding of lines that were selected for high versus low breeding values for male courtship rate (two high lines, two unselected control lines, two low lines; generations 'S1 to S4'; see [2,3]). The birds used in experiments 6, 7, and 8 were from the generations F2, F3, and S3, respectively.

Allocation of color bands in experiment 6

One purpose of experiment 6 was to test the hypothesis that individual birds would develop a preference for the traits of their partner. Hence, birds were randomly allocated striped color bands at the beginning of the experiment, and after 14 days when most pairs had formed, each pair was allocated a pre-defined combination of colors (through a random process). The allocation was done in such a way that identical combinations were used in two separate rooms in which 'parallel worlds' were created. For instance, aviary 1 in room 1 would contain five pairs with the same color band combinations as in aviary 1 in room 2. As usual (see Methods), each of the six colors was used only once per sex and aviary, and pairs would typically wear different colors (e.g. white-banded male with orange-banded female). At the end of the experiment (after all young were reared to independence), the birds of one sex were swapped among the 'parallel worlds', such that the white-banded male, who had just lost his orange-banded partner, would be placed in an aviary with one available (unfamiliar) orange-banded female who also just lost her white-banded partner. Using such 'parallel worlds' allowed to test the hypothesis that birds used the color traits of their previous partner as a search image for a new partner. This hypothesis predicted specific pairings (e.g. white-banded male with orange-banded female). Our experiment revealed no support for this hypothesis: pair formation after swapping between parallel worlds occurred randomly with regard to the band color of the previous partner. For the

purpose of the present study, allocating specific color bands to pairs that had formed previously without color bands (e.g. white with orange, used two times among the six aviaries) should not have induced any bias, especially because we ensured that matched pairs (with similar color bands in the two rooms) were not matched for the timing of pair formation. In other words, the color combinations used in aviary 1 in room 1 and in aviary 1 in room 2 were randomized across pairing order, independently for the two aviaries.

Table S1. Summary of publications (n = 39) [4-42] from studies in which male zebra finches were fitted with red versus green color bands. Studies were classified as (1) providing support ('Result' = 1, n = 23) for the hypothesis that red-banded males are in some way doing 'better' than green-banded males, (2) providing partial support ('Result' = 0.5, n = 8) defined as showing at least some significant effect of color bands, or (3) no support ('Result = 0, n = 8) defined as showing no significant effects of color bands. Indicated are the reference, the main focus of the study, the number of females and males that were studied (sample size might deviate slightly depending on the trait studied and on exclusion of individuals), and the result. Note that samples were numbers of tested (choosing) individuals during the mate preference experiments. For studies in which many traits were investigated (e.g. Zann 1994), the biggest sample size was used here. A list of complete references was added at the end of this supplement.

Year	Result	Study trait	N females	N males	Author
1981	1	sex ratio	30	30	Burley
1982	1	mate preference	17	38	Burley et al
1985	1	mate preference	18	17	Burley
1985	1	mortality	24	24	Burley
1986	1	reproductive success	24	24	Burley
1986	1	sex ratio	24	24	Burley
1986	1	mate preference	15	18	Burley
1987	0	sexual trait		36	Ratcliffe & Boag
1988	1	mate preference	24		Burley
1988	1	parental care	16	31	Burley
1994	1	sexual trait		31	Burley et al
1994	0.5	life history	279	194	Zann
1996	1	sexual trait		30	Burley et al
1996	1	sexual trait		36	Swaddle
1997	1	competition		32	Cuthill et al
1997	1	mate preference	24		Hunt et al
1998	0	mate preference	10		Jennions
1999	1	maternal effect	12	24	Gil et al
1999	0	social leaning		36	Peason et al
1999	0.5	sexual trait	10		Waas & Wordsworth
2002	1	social leaning	7	7	Benskin et al
2003	0	sex ratio	20	20	Zann & Runciman
2004	0	mate preference	36		Nakagawa & Waas
2004	1	mate preference	15		Burley & Foster
2004	0.5	life history		50	Rutstein et al
2005	0.5	parental are		35	Gorman et al
2005	0.5	sex ratio		70	Rutstein et al
2006	1	maternal effect		36	Gilbert et al
2006	1	sexual trait		52	Gleeson

2006	0.5	maternal effect		10	Williamson et al
2006	1	mate preference	12		Burley
2006	1	mate preference	16		Burley & Foster
2010	1	sexual trait		58	Pariser et al
2010	0	sexual trait		30	Schuett & Dall
2012	1	life history		70	Gilbert et al
2012	0	sexual trait		153	Seguin& Forstmeier
2014	0	social leaning		60	Mora & Forstmeier
2016	0.5	parental are		76	Arnold et al
2017	0.5	mate preference	71		Song et al

61

Table S2. Mean relative fitness of female zebra finches with different color bands. Data for each experiment are shown, and the population and sample size (N) are indicated.

Experiment	Population	Color	Relative fitness	N
1	Melbourne	white	1.53	5
1	Melbourne	black	1.51	5
1	Melbourne	yellow	0.94	5
1	Melbourne	light green	0.80	5
1	Melbourne	blue	0.70	2
1	Melbourne	light blue	0.60	2
1	Melbourne	red	0.50	5
2	Bielefeld	light green	2.03	5
2	Bielefeld	light blue	1.85	3
2	Bielefeld	white	1.31	5
2	Bielefeld	black	0.70	5
2	Bielefeld	red	0.66	5
2	Bielefeld	yellow	0.31	5
2	Bielefeld	blue	0.14	3
3	Bielefeld	blue	1.08	16
3	Bielefeld	light blue	1.06	16
3	Bielefeld	white	1.00	17
3	Bielefeld	orange	0.97	17
3	Bielefeld	yellow	0.95	17
3	Bielefeld	black	0.94	17
4	Bielefeld	blue	1.62	6
4	Bielefeld	black	1.22	6
4	Bielefeld	orange	1.01	6
4	Bielefeld	white	0.82	6
4	Bielefeld	yellow	0.77	6
4	Bielefeld	light blue	0.56	6
5	Krakow	yellow	1.72	8
5	Krakow	light blue	1.11	4
5	Krakow	red	1.07	8
5	Krakow	white	0.95	8
5	Krakow	blue	0.92	4
5	Krakow	black	0.81	8
5	Krakow	light green	0.44	8
6(1)	Seewiesen	light green-black	1.35	6
6(1)	Seewiesen	red-black	1.16	6
6(1)	Seewiesen	red-white	0.95	6
6(1)	Seewiesen	white-black	0.90	6
6(1)	Seewiesen	yellow-black	0.86	6
6(1)	Seewiesen	light green-white	0.77	6
6(2)	Seewiesen	pink	1.47	6
6(2)	Seewiesen	white	1.36	6

6(2)	Seewiesen	black	1.21	6
6(2)	Seewiesen	blue	0.90	6
6(2)	Seewiesen	yellow	0.59	6
6(2)	Seewiesen	orange	0.47	6
7	Seewiesen	orange	1.28	6
7	Seewiesen	yellow	1.08	6
7	Seewiesen	blue	1.05	6
7	Seewiesen	black	1.04	6
7	Seewiesen	white	0.96	6
7	Seewiesen	pink	0.57	6
8	Seewiesen	black	1.31	15
8	Seewiesen	pink	1.19	15
8	Seewiesen	light green	1.15	15
8	Seewiesen	light blue	0.88	15
8	Seewiesen	orange	0.83	15
8	Seewiesen	red	0.64	15

Table S3. Mean relative fitness of male zebra finches with different color bands. Data for each experiment are shown, and the population and sample size (N) are indicated.

Experiment	Population	Color	Relative fitness	N
1	Melbourne	light green	1.36	5
1	Melbourne	yellow	1.25	5
1	Melbourne	red	1.09	5
1	Melbourne	blue	1.09	3
1	Melbourne	white	1.01	5
1	Melbourne	light blue	0.92	3
1	Melbourne	black	0.29	5
2	Bielefeld	white	2.12	5
2	Bielefeld	light blue	1.83	2
2	Bielefeld	red	0.82	5
2	Bielefeld	black	0.74	5
2	Bielefeld	light green	0.73	5
2	Bielefeld	blue	0.64	2
2	Bielefeld	yellow	0.39	5
3	Bielefeld	orange	1.25	17
3	Bielefeld	light blue	1.06	16
3	Bielefeld	yellow	1.02	17
3	Bielefeld	blue	0.92	16
3	Bielefeld	black	0.91	17
3	Bielefeld	white	0.84	17
4	Bielefeld	yellow	1.33	6
4	Bielefeld	light blue	1.30	6
4	Bielefeld	black	1.25	6
4	Bielefeld	white	0.79	6
4	Bielefeld	orange	0.73	6
4	Bielefeld	blue	0.60	6
5	Krakow	yellow	1.22	8
5	Krakow	light green	1.20	8
5	Krakow	blue	1.05	4
5	Krakow	light blue	0.94	4
5	Krakow	black	0.93	8
5	Krakow	red	0.84	8
5	Krakow	white	0.81	8
6(1)	Seewiesen	red-white	1.64	6
6(1)	Seewiesen	yellow-black	1.29	6
6(1)	Seewiesen	red-black	1.06	6

6(1)	Seewiesen	white-black	0.80	6
6(1)	Seewiesen	light green-black	0.75	6
6(1)	Seewiesen	light green-white	0.46	6
6(2)	Seewiesen	black	1.26	6
6(2)	Seewiesen	yellow	1.19	6
6(2)	Seewiesen	white	1.02	6
6(2)	Seewiesen	blue	0.93	6
6(2)	Seewiesen	orange	0.92	6
6(2)	Seewiesen	pink	0.67	6
7	Seewiesen	pink	1.21	6
7	Seewiesen	blue	1.14	6
7	Seewiesen	white	1.09	6
7	Seewiesen	black	0.89	6
7	Seewiesen	yellow	0.88	6
7	Seewiesen	orange	0.79	6
8	Seewiesen	pink	1.25	15
8	Seewiesen	black	1.09	15
8	Seewiesen	orange	1.08	15
8	Seewiesen	light blue	0.97	15
8	Seewiesen	red	0.80	15
8	Seewiesen	light green	0.80	15

70

71

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

	estimate (β±SE)	t	*p*
random effects:			
Aviary (n = 39)	0.000		
Band color (n = 9)	0.000		
Experiment (n = 5)	0.043		
Population (n = 4)	0.000		
Residual	1.064		
fixed effects:			
Intercept	1.378±0.119	11.6	-
Adult sex ratio	1.092±1.212	0.90	0.37
Duration of breeding season (d)	0.020±0.014	1.40	0.16
Inbreeding coefficient	-3.872±0.986	-3.93	<0.0001

Table S5. Linear mixed model explaining the 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). For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

	estimate (β±SE)	t	*p*
random effects:			
Aviary (n = 39)	0.000		
Band color (n = 9)	0.000		
Experiment (n = 5)	0.000		
Population (n = 4)	0.013		
Residual	1.279		
fixed effects:			
Intercept	1.315±0.100	13.2	-
Adult sex ratio	-3.192±1.323	-2.41	0.02
Duration of breeding season	0.015±0.010	1.53	0.13
Inbreeding coefficient	-3.102±1.029	-3.01	0.003
Red versus green band	-0.304±0.279	-1.09	0.28

Table S6. Linear mixed model of female fitness using the 'attractiveness rank' as a fixed effect (defined for 6 colors only). This model corresponds to Figure 4 top left (female, version1). Striped color bands containing green or red from experiment 6(1) were lumped with the categories 'green' and 'red'. For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

	estimate (β±SE)	t	*p*
random effects:			
Female ID (n = 253)	0.422		
Aviary (n = 68)	0.000		
Band color (n = 6)	0.000		
Experiment (n = 8)	0.008		
Population (n = 4)	0.000		
Residual	0.682		
fixed effects:			
Intercept	1.523±0.076	20.16	-
Adult sex ratio	0.857±1.603	0.54	0.59
Duration of breeding season	0.002±0.011	0.20	0.84
Inbreeding coefficient	-3.064±0.813	-3.77	0.0002
Attractiveness rank	0.043±0.162	0.26	0.79

Table S7. Linear mixed model of female fitness using the 'attractiveness rank' as a fixed effect (defined for 6 colors only). This model corresponds to Figure 4 top right (female, version2). The analysis is based on uniformly colored bands from experiment 6(2). For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

	estimate (β±SE)	t	*p*
random effects:			
Female ID (n = 247)	0.429		
Aviary (n = 68)	0.000		
Band color (n = 6)	0.012		
Experiment (n = 8)	0.000		
Population (n = 4)	0.000		
Residual	0.687		
fixed effects:			
Intercept	1.492±0.082	18.2	-
Adult sex ratio	0.918±1.613	0.57	0.57
Duration of breeding season	0.002±0.011	0.17	0.86
Inbreeding coefficient	-2.983±0.780	-3.82	0.0001
Attractiveness rank	0.098±0.198	0.49	0.62

Table S8. Linear mixed model of male fitness using the 'attractiveness rank' as a fixed effect (defined for 6 colors only). This model corresponds to Figure 4 bottom left (male, version1). Striped color bands containing green or red from experiment 6(1) were lumped with the categories 'green' and 'red'. For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

	estimate (β±SE)	t	*p*
random effects:			
Male ID (n = 253)	0.330		
Aviary (n = 68)	0.000		
Band color (n = 6)	0.009		
Experiment (n = 8)	0.017		
Population (n = 4)	0.000		
Residual	0.876		
fixed effects:			
Intercept	1.500±0.097	15.4	-
Adult sex ratio	-3.219±1.682	-1.91	0.06
Duration of breeding season	-0.002±0.013	-0.13	0.90
Inbreeding coefficient	-4.107±0.975	-4.21	<0.0001
Attractiveness rank	-0.060±0.226	-0.27	0.79

Table S9. Linear mixed model of male fitness using the 'attractiveness rank' as a fixed effect (defined for 6 colors only). This model corresponds to Figure 4 bottom right (male, version2). The analysis is based on uniformly colored bands from experiment 6(2). For random effects, the size of the variance component is shown. All fixed effects were mean-centered.

	estimate (β±SE)	t	*p*
random effects:			
Male ID (n = 247)	0.359		
Aviary (n = 68)	0.000		
Band color (n = 6)	0.000		
Experiment (n = 8)	0.023		
Population (n = 4)	0.000		
Residual	0.863		
fixed effects:			
Intercept	1.460±0.095	15.4	-
Adult sex ratio	-3.292±1.692	-1.95	0.05
Duration of breeding season	-0.001±0.013	-0.04	0.96
Inbreeding coefficient	-3.899±0.991	-3.94	<0.0001
Attractiveness rank	-0.266±0.220	-1.02	0.23

Additional References

1. Forstmeier W, Segelbacher G, Mueller JC, Kempenaers B. Genetic variation and differentiation in captive and wild zebra finches (Taeniopygia guttata). Mol Ecol. 2007;16(19):4039-50. doi: 10.1111/j.1365-294X.2007.03444.x. PubMed PMID: WOS:000249829500007.
2. Mathot KJ, Martin K, Kempenaers B, Forstmeier W. Basal metabolic rate can evolve independently of morphological and behavioural traits. Heredity. 2013;111(3):175-81. doi: 10.1038/hdy.2013.35. PubMed PMID: ISI:000323209400001.
3. Wang D, Kempenaers N, Kempenaers B, Forstmeier W. Male zebra finches have limited ability to identify high-fecundity females. Behav Ecol. 2017;28(3):784-92.
4. Burley N. Sex-Ratio Manipulation and Selection for Attractiveness. Science. 1981;211(4483):721-2. doi: DOI 10.1126/science.211.4483.721. PubMed PMID: WOS:A1981LA65300029.
5. Burley N, Krantzberg G, Radman P. Influence of Color-Banding on the Conspecific Preferences of Zebra Finches. Anim Behav. 1982;30(May):444-55. doi: Doi 10.1016/S0003-3472(82)80055-9. PubMed PMID: WOS:A1982NQ92800014.
6. Burley N. The organization of behavior and the evolution of sexually selected traits. Avian monogamy. 1985;37:22-44.
7. Burley N. Leg-Band Color and Mortality Patterns in Captive Breeding Populations of Zebra Finches. Auk. 1985;102(3):647-51. PubMed PMID: WOS:A1985ANS9200028.
8. Burley N. Sexual selection for aesthetic traits in species with biparental care. The American Naturalist. 1986;127(4):415-45.
9. Burley N. Sex-Ratio Manipulation in Color-Banded Populations of Zebra Finches. Evolution. 1986;40(6):1191-206. doi: Doi 10.2307/2408947. PubMed PMID: WOS:A1986F052900008.
10. Burley N. Comparison of the band-colour preferences of two species of estrildid finches. Anim Behav. 1986;34(6):1732-41.
11. Ratcliffe LM, Boag PT. Effects of Color Bands on Male Competition and Sexual Attractiveness in Zebra Finches (Poephila-Guttata). Can J Zool. 1987;65(2):333-8. PubMed PMID: WOS:A1987G472400018.
12. Burley N. Wild Zebra Finches Have Band-Color Preferences. Anim Behav. 1988;36:1235-7. doi: Doi 10.1016/S0003-3472(88)80085-X. PubMed PMID: WOS:A1988P531400032.
13. Burley N. The differential-allocation hypothesis: an experimental test. The American Naturalist. 1988;132(5):611-28.
14. Burley NT, Enstrom DA, Chitwood L. Extra-Pair Relations in Zebra Finches - Differential Male Success Results from Female Tactics. Anim Behav. 1994;48(5):1031-41. doi: DOI 10.1006/anbe.1994.1336. PubMed PMID: WOS:A1994PQ35100005.
15. Zann R. Effects of Band Color on Survivorship, Body Condition and Reproductive Effort of Free-Living Australian Zebra Finches. Auk. 1994;111(1):131-42. PubMed PMID: WOS:A1994NL77800014.
16. Burley NT, Parker PG, Lundy K. Sexual selection and extrapair fertilization in a socially monogamous passerine, the zebra finch (Taeniopygia guttata). Behav Ecol. 1996;7(2):218-26. doi: DOI 10.1093/beheco/7.2.218. PubMed PMID: WOS:A1996UK92900018.
17. Swaddle JP. Reproductive success and symmetry in zebra finches. Anim Behav. 1996;51:203-10. doi: DOI 10.1006/anbe.1996.0017. PubMed PMID: WOS:A1996TR87400017.
18. Cuthill IC, Hunt S, Cleary C, Clark C. Colour bands, dominance, and body mass regulation in male zebra finches (Taeniopygia guttata). P Roy Soc B-Biol Sci. 1997;264(1384):1093-9. PubMed PMID: WOS:A1997XP13100022.
19. Hunt S, Cuthill IC, Swaddle JP, Bennett ATD. Ultraviolet vision and band-colour preferences in female zebra finches, Taeniopygia guttata. Anim Behav. 1997;54:1383-92. doi: DOI 10.1006/anbe.1997.0540. PubMed PMID: WOS:000071356400005.
20. Jennions MD. The effect of leg band symmetry on female-male association in zebra finches. Anim Behav. 1998;55:61-7. doi: DOI 10.1006/anbe.1997.0579. PubMed PMID: WOS:000071789700007.

21. Gil D, Graves J, Hazon N, Wells A. Male attractiveness and differential testosterone investment in zebra finch eggs. Science. 1999;286(5437):126-8. doi: DOI 10.1126/science.286.5437.126. PubMed PMID: WOS:000082907400051.
22. Pearson FD, Mann NI, Slater PJB. Does leg-ring colour affect song tutor choice in zebra finches? Anim Behav. 1999;57:173-80. doi: DOI 10.1006/anbe.1998.0924. PubMed PMID: WOS:000078341700019.
23. Waas JR, Wordsworth AF. Female zebra finches prefer symmetrically banded males, but only during interactive mate choice tests. Anim Behav. 1999;57:1113-9. doi: DOI 10.1006/anbe.1998.1055. PubMed PMID: WOS:000080285600015.
24. Benskin CMWH, Mann NI, Lachlan RF, Slater PJB. Social learning directs feeding preferences in the zebra finch, Taeniopygia guttata. Anim Behav. 2002;64:823-8. doi: 10.1006/anbe.2002.2005. PubMed PMID: WOS:000179701200016.
25. Zann R, Runciman D. Primary sex ratios in zebra finches: no evidence for adaptive manipulation in wild and semi-domesticated populations. Behav Ecol Sociobiol. 2003;54(3):294-302. doi: 10.1007/s00265-003-0632-1. PubMed PMID: WOS:000184754900012.
26. Burley NT, Foster VS. Digit ratio varies with sex, egg order and strength of mate preference in zebra finches. P Roy Soc B-Biol Sci. 2004;271(1536):239-44. doi: 10.1098/rspb.2003.2562. PubMed PMID: WOS:000188694400004.
27. Nakagawa S, Waas JR. The effect of acoustic and visual priming stimuli on the reproductive behaviour of female zebra finches, Taeniopygia guttata. acta ethologica. 2004;7(1):43-9.
28. Rutstein AN, Gilbert L, Slater PJB, Graves JA. Mate attractiveness and primary resource allocation in the zebra finch. Anim Behav. 2004;68:1087-94. doi: DOI 10.1016/j.anbehav.2004.02.011. PubMed PMID: WOS:000225064400013.
29. Gorman HE, Arnold KE, Nager RG. Incubation effort in relation to male attractiveness in zebra finches Taeniopygia guttata. J Avian Biol. 2005;36(5):413-20. PubMed PMID: WOS:000232567600008.
30. Rutstein AN, Gorman HE, Arnold KE, Gilbert L, Orr KJ, Adam A, et al. Sex allocation in response to paternal attractiveness in the zebra finch. Behav Ecol. 2005;16(4):763-9. doi: 10.1093/beheco/ari052. PubMed PMID: WOS:000229856900012.
31. Gilbert L, Williamson KA, Hazon N, Graves JA. Maternal effects due to male attractiveness affect offspring development in the zebra finch. P Roy Soc B-Biol Sci. 2006;273(1595):1765-71. doi: 10.1098/rspb.2006.3520. PubMed PMID: WOS:000238946100007.
32. Gleeson DJ. Context-dependent effect of social environment on immune response and sexual signalling in male zebra finches. Aust J Zool. 2006;54(6):375-9. doi: 10.1071/Zo06001. PubMed PMID: WOS:000243335900001.
33. Williamson KA, Surai PF, Graves JA. Yolk antioxidants and mate attractiveness in the Zebra Finch. Funct Ecol. 2006;20(2):354-9. doi: 10.1111/j.1365-2435.2006.01087.x. PubMed PMID: WOS:000237517400022.
34. Burley NT, Foster VS. Variation in female choice of mates: condition influences selectivity. Anim Behav. 2006;72:713-9. doi: 10.1016/j.anbehav.2006.01.017. PubMed PMID: WOS:000240730300025.
35. Burley NT. An eye for detail: Selective sexual imprinting in zebra finches. Evolution. 2006;60(5):1076-85. PubMed PMID: WOS:000238181700017.
36. Pariser EC, Mariette MM, Griffith SC. Artificial ornaments manipulate intrinsic male quality in wild-caught zebra finches (Taeniopygia guttata). Behav Ecol. 2010;21(2):264-9. doi: 10.1093/beheco/arp185. PubMed PMID: WOS:000274483400009.
37. Schuett W, Dall SRX. Appearance, "state," and behavior in male zebra finches, Taeniopygia guttata. J Ethol. 2010;28(2):273-86. doi: 10.1007/s10164-009-0181-9. PubMed PMID: WOS:000277469300009.

212 38. Gilbert L, Williamson KA, Graves JA. Male attractiveness regulates daughter fecundity non-
213 genetically via maternal investment. P Roy Soc B-Biol Sci. 2012;279(1728):523-8. doi:
214 10.1098/rspb.2011.0962. PubMed PMID: WOS:000298661700014.
215 39. Seguin A, Forstmeier W. No Band Color Effects on Male Courtship Rate or Body Mass in the
216 Zebra Finch: Four Experiments and a Meta-Analysis. Plos One. 2012;7(6). doi: ARTN
217 e3778510.1371/journal.pone.0037785. PubMed PMID: WOS:000305351700008.
218 40. Mora AR, Forstmeier W. The Importance of Validating Experimental Setups: Lessons from Studies
219 of Food Choice Copying in Zebra Finches. Ethology. 2014;120(9):913-22. doi: 10.1111/eth.12263.
220 PubMed PMID: WOS:000339567300008.
221 41. Arnold KE, Gilbert L, Gorman HE, Griffiths KJ, Adam A, Nager RG. Paternal attractiveness and the
222 effects of differential allocation of parental investment. Anim Behav. 2016;113:69-78. doi:
223 10.1016/j.anbehav.2015.12.013. PubMed PMID: WOS:000371257300011.
224 42. Song Z, Liu Y, Booksmythe I, Ding C. Effects of individual-based preferences for colour-banded
225 mates on sex allocation in zebra finches. Behav Ecol. 2017.
226

227

Chapter 4: Scrutinizing assortative mating in birds

Short title: Assoatative mating in birds

Abstract: Pair bonds often form between individuals that resemble one another. Such assortative mating appears to be widespread not only in humans but also throughout the animal kingdom. Yet it remains usually unclear whether assortative mating arises primarily from mate choice ('like attracts like'), from spatial or temporal separation, or from observer, reporting, publication and search bias. Here, we reveal how compelling meta-analytical evidence for size-assortative mating in birds (r = 0.201 ± 0.022 SE, 58 species, 15,971 pairs) vanishes gradually with increased control of confounding factors. Specifically, the effect size decreased to half when we estimated assortative mating from unpublished data (free of reporting and publication bias) of nine long-term field studies (r = 0.106 ± 0.048 SE, eight species, 16,611 pairs) and assortative mating nearly disappeared (to around r = 0.018) when both partners were measured by independent observers or separate in space and time. Finally, we found no evidence for assortative mating in a direct experimental test for mutual mate choice in captive populations of zebra finches (r = -0.003 ± 0.141 SE, 1,414 pairs). These results highlight the importance of unpublished data in generating unbiased meta-analytical conclusions, and suggest that the apparent ubiquity of assortative mating reported in the literature is overestimated and may typically not be driven by mate choice.

Prepared as: Daiping Wang, Wolfgang Forstmeier, Mihai Valcu, Niels Dingemanse, Martin Bulla, Christiaan Both, Renee Duckworth, Lynna Marie Kiere, Patrik Karell, Tomáš Albrecht, Bart Kempenaers: Scrutinizing assortative mating in birds

Scrutinizing assortative mating in birds

Daiping Wang[1], Wolfgang Forstmeier[1] Mihai Valcu[1], Niels Dingemanse[2], Martin Bulla[1,3,4], Christiaan Both[5], Renée A. Duckworth[6], Lynna Marie Kiere [7,8], Patrik Karell[9], Tomáš Albrecht[10], Bart Kempenaers[1]

1. Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany

2. Behavioural Ecology, Department of Biology, Ludwig Maximilians University of Munich, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany

3. NIOZ Royal Netherlands Institute for Sea Research, Department of Coastal Systems and Utrecht University, P.O. Box 59, 1790 AB Den Burg, The Netherlands

4. Department of Ecology, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague 6 - Suchdol, Czech Republic

5. Conservation Ecology Group, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

6. Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

7. Center for the Study of Biodiversity and Conservation (CIByC), Autonomous University of the State of Morelos, Cuernavaca Morelos, Mexico

8. Institute of Ecology, Department of Evolutionary Ecology, Universidad Nacional Autónoma de México, Mexico City, Distrito Federal, Mexico

9. Bioeconomy Research Team, Novia University of Applied Sciences, Raseborgsvägen 9, FI-10600 Ekenäs, Finland

10. Institute of Vertebrate Biology of the Czech Academy of Sciences, 603 65, Brno, and Faculty of Science, Charles University, 128 44 Prague, Czech Republic

Short title: Assortative mating in birds

Keywords: mate choice, effect size, measurement error, meta-analysis, publication bias, observer effect, spatial and temporal autocorrelation

Figures & Tables: 2 figures, 1 table

Supplementary materials: 1 figure, 7 tables

Address for correspondence:

Wolfgang Forstmeier, Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Str. 7, 82319 Seewiesen, Germany, Phone: 0049-8157-932346

Email: forstmeier@orn.mpg.de

Abstract

Pair bonds often form between individuals that resemble one another. Such assortative mating appears to be widespread not only in humans but also throughout the animal kingdom. Yet it remains usually unclear whether assortative mating arises primarily from mate choice ('like attracts like'), from spatial or temporal separation, or from observer, reporting, publication and search bias. Here, we reveal how compelling meta-analytical evidence for size-assortative mating in birds (r = 0.201 ± 0.022 SE, 58 species, 15,971 pairs) vanishes gradually with increased control of confounding factors. Specifically, the effect size decreased to half when we estimated assortative mating from unpublished data (free of reporting and publication bias) of nine long-term field studies (r = 0.106 ± 0.048 SE, eight species, 16,611 pairs) and assortative mating nearly disappeared (to around r = 0.018) when both partners were measured by independent observers or separate in space and time. Finally, we found no evidence for assortative mating in a direct experimental test for mutual mate choice in captive populations of zebra finches (r = -0.003 ± 0.141 SE, 1,414 pairs). These results highlight the importance of unpublished data in generating unbiased meta-analytical conclusions, and suggest that the apparent ubiquity of assortative mating reported in the literature is overestimated and may typically not be driven by mate choice.

Key words: mate choice, effect size, measurement error, meta-analysis, publication bias, observer effect, spatial and temporal autocorrelation

Introduction

Members of a pair often resemble each other. For instance, in humans partners have similar political attitudes[1,2] (Alford et al. 2011; Klofstad et al. 2013), level of education[3,4] (Domingue et al. 2014; Robinson et al. 2017), and body height[4-6] (Tenesa et al. 2016; Robinson et al. 2017; Stulp et al. 2017). Assortative mating appears to be pervasive across all animal taxa and across all phenotypic traits that have been investigated (for a recent meta-analysis see [7] Jiang et al. 2013). However, in most cases, the underlying processes that lead to mate similarity remain unclear.

Similarity of pair members, quantified as the strength of the correlation between their trait values, may arise via three biological mechanisms. (1) Mate choice. One or both sexes may prefer phenotypes similar to their own ('like attracts like'). This may lead to the more frequent formation and enhanced stability of assortative pair bonds. (2) Spatial and temporal autocorrelation. Individuals with different phenotypes may be separated in space and time, such that at the population level even random mating would lead to partner similarity ('like meets like'). For instance, in high-quality habitats individuals may grow larger than in poor habitats. If individuals from different habitats are less likely to meet ('non-panmixis'), e.g. because of reduced mobility[8] (Rolan-Alvaret et al. 2015), a population-wide pattern of assortative mating may arises in the absence of choice for an assortative partner. Similarly, in migratory species, individuals that resemble each other in particular traits may have a higher probability to form a pair simply because they arrive at the breeding grounds closer in time (e.g. older individuals might arrive earlier, leading to assortative mating for age[9] Village 1985). (3) Phenotypic changes over time. Females and males may mate randomly for a certain phenotype, but become similar to their partner over time ('become alike')[10] (Anderson et al. 2003). For instance, in humans a positive correlation in body mass between couples may arise because they share the same food[11,12] (Price and Vandenberg 1980; Feunekes et al. 1997).

The three biological mechanisms can act together and their relative importance may be difficult to tease apart. Assortative mating is often investigated with a focus on mate choice[13,14] (Houtman and Falls 1994; García-Navas et al. 2009). In that case, the other two mechanisms ('like meets like' and 'become alike') may confound the results. When individuals are separated in space or time, evidence for the role of mate choice requires knowledge about the potential partners available during pair formation (who encountered whom). Note that separation in space or time according to certain phenotypic traits might already be part of the mate choice process. In this case, an experimental approach would be needed to provide evidence for

preference for a similar partner. To assess the importance of individuals 'becoming alike', phenotypic measurements need to be taken at the time of mating and again later on, or the duration of the pair bond needs to be included in the analysis. In field studies such information may be difficult to obtain.

Besides the influence of biological processes, estimates of the strength of assortative mating can also be confounded by several methodological issues. (1) Observer bias. Data-sets often consist of measurements from multiple observers and taken over longer periods. Trait correlations between pair members may then arise when pair members are measured by the same observer and on the same day, because of consistent between-observer differences in measurements[15] (Cunningham et al. 1999) and because observers may (unconsciously) change their measuring technique over time. (2) Reporting bias. Estimates found in the literature will be inflated when statistically significant estimates are more likely reported than non-significant ones[16,17] (Greenwald 1975; Open-Science-Collaboration 2015). (3) Search bias. If a meta-analysis is based on a literature search with key-words like "assortative", the strength of assortative mating may be overestimated, because null results may be less likely mentioned in the abstract of a publication[18] (Kulshrestha et al. 2017) and hence such searches may preferentially yield a subset of studies that have detected significant assortative mating. Similarly, when screening relevant publications, taking estimates from related studies that are being cited may also discriminate against null findings, because studies with null findings tend to get cited less often than studies with significant and hence typically larger effects[19] (Ferguson & Heene 2012) .

Here, we quantify the strength of assortative mating and assess how estimates change with increasing control for the confounding factors discussed above. We then propose ways to minimize confounding effects if the aim is to investigate assortative mating due to mate choice. For practical reasons (data availability), we focus primarily on assortative mating for size in birds, but our approach is relevant for most phenotypic traits.

First, we compare published estimates of the strength of assortative mating with estimates from unpublished data from nine long-term field studies. This allows assessing the effect of search and reporting bias, which should only affect the published dataset. Second, we use the unpublished dataset to explore the effects of observer bias and spatial and temporal independence of the measurements on the estimates of assortative mating. Finally, we present an analysis of experimental data from studies of assortative mating in captive zebra finches *Taeniopygia guttata*[20,21] (Ihle et al. 2015, Wang et al. 2017). In these experiments, we took

standardized measurements of all birds before randomly allocating them to experimental aviaries. In this way, we can estimate the strength of assortative mating among individuals that encountered each other, excluding all known confounding factors.

Results

Assortative mating: all traits - published literature

Overall, the published literature showed considerable evidence for positive assortative mating across all trait categories (ranging from r = 0.198 to 0.409; none of the 95% CI overlap zero; Figure 1a, Table S1). The two random effects 'Study' and 'Species' explained only 12% and 2% of the variance, respectively. This means that levels of assortative mating were slightly repeatable across traits within studies, but not between studies of the same species. Compared to other traits, assortative mating for body size was the weakest (r = 0.198), but it was also the most frequently studied trait (57% of all estimates).

Assortative mating for size: unpublished field studies

The unpublished data from nine long-term field studies also showed a clear, yet weaker tendency for positive assortative mating by size, but the magnitude depended on how the data were analysed (Figure S1).

When repeatedly measured individuals were represented by a randomly selected single measure (i.e. model of 'random alignment' of male to female measure) the strength of assortment was weak (r = 0.070, based on 16,545 pair-trait combinations, Figure S1a, Table S2). When average measures per individual were used ('average model'), estimates of assortment were only slightly higher (r = 0.082, n = 16,545, Figure S1b, Table S3). Finally, when using the male and female measures taken closest to the presumed time of pair formation ('nearest model'), the estimate of assortative mating was highest (r = 0.102, n = 16,611, Figure S1c, Table S4). The estimates from the 'nearest' model were significantly higher than those from the 'random alignment' model (paired t-test for 32 species-traits, $t_{31} = 4.50$, p = 0.0001), and from the 'average' model ($t_{31} = 3.32$, p = 0.002).

Effects of observer, time, and space on estimates of assortative mating

In the unpublished dataset, levels of apparent assortative mating were significantly higher when measurements on the two members of a pair had been taken by the same observer (r = 0.075 ± 0.021, t = 3.45, p = 0.0006, n = 22 estimates, 34,672 pair-trait combinations) than when

measurements came from different observers (r = 0.023 ± 0.022, t = 1.01, p = 0.3, n = 22 estimates, 24,771 pair-trait combinations; $t_{observer}$ = 2.48, p = 0.01; Figure 1b, Table S5: model 4).

Similarly, estimates for assortative mating for size were significantly higher when measurements on the two members of a pair had been taken within 30 days of each other (r = 0.110 ± 0.016, t = 6.92, p < 0.0001, n = 32 estimates, 51,995 pair-trait combinations) than when the partners had been measured more than 30 days apart (r = 0.014 ± 0.014, t = 0.98, p = 0.3, n = 31 estimates, 20,729 pair-trait combinations; (t_{time} = 5.7, p < 0.0001; Figure 1b, Table S5: model 5).

Finally, estimates of size-assortative mating were significantly higher when partners had been measured at the same site (0.073 ± 0.014, t = 5.30, p < 0.0001, n = 32 estimates, 26,542 pair-trait combinations) than when they were measured at different sites (0.017 ± 0.014, t = 1.28, p = 0.2, n = 32 estimates, 44,112 pair-trait combinations; $t_{location}$ = 3.90, p < 0.0001; Figure 1b, Table S5: model 6).

Assortative mating for size: experimental study

Data from the five experiments on zebra finches showed an overall size-assortative mating close to zero (r = -0.020, weighted mean of 13 estimates, n = 1,414 pair-trait combinations; Table S6). Note that the statistical power for detecting an effect of r = 0.20 was >0.99 for each of the three size phenotypes.

Effect of data source on estimates of size-assortative mating

The estimates of the strength of assortative mating decreased with increasing control for confounding factors from published through unpublished to experimental data (Figure 1c, Table S7).

Effect of sample size on estimates of assortative mating strength

Figure 2 shows the individual correlation coefficients (from all four data sources) in relation to the sample size on which they are based. We found limited evidence for asymmetry in the funnel plot for the literature data ('WoS Search Search' and 'Cited studies' jointly: t = -1.73, p = 0.084, n = 357), suggesting only a modest decline of estimates of assortative mating from low to high sample size. For high sample sizes, the correlation coefficients from the literature and from the 'Unpublished data' are similar (compare the red and blue regression lines in Figure 2).

Discussion

Our meta-analysis of published estimates shows clear evidence of positive assortative mating in birds across different phenotypes (Figure 1a). This is consistent with a recent meta-analysis of assortative mating across the whole animal kingdom[7] (Jiang et al. 2013). However, this study also suggests that these results cannot be taken as evidence for mate choice for a similar partner ('like attracts like'). First, the underlying causes for positive assortment of mates usually remain unclear, and various confounding factors may have inflated the estimates of assortative mating. Our study reveals that seemingly robust effects may largely disappear when controlling for multiple sources of bias (Figure 1b, c), and hence question the ubiquity and importance of assortative mating. In the following, we discuss the effects of each confounding factor and where possible suggest ways to avoid the bias. We also discuss the current evidence for assortative mating for size and other traits in relation to mate choice or other processes.

Evidence for search and reporting bias

Neither our meta-analysis of size-assortative mating in birds (Figure 2), nor the recent analysis across the animal kingdom[7] (Jiang et al. 2013) found strong evidence for publication bias as indicated by significant asymmetry in the funnel plot (we found a non-significant trend in the expected direction, $p = 0.084$). This could indicate either that publication bias is limited, or that tests for asymmetry in the funnel plot are inefficient in detecting it[22] (Tang and Liu 2000). Some bias is expected, because most studies emphasize positive findings rather than null results[23] (Fanelli 2010) and because incomplete reporting of non-significant outcomes is widespread[24] (Kittelman et al. 2018).

A different way to test for these biases is to contrast published with unpublished estimates (Figure 1c). So far, only few meta-analyses have included such comparison, but those who did found that published effect sizes were larger than unpublished ones (e.g.[25-27]) (Coltman & Slate 2003, Wang et al. 2018, Sanchez-Tojar et al. 2018). The use of unpublished data or of 'grey literature' has been criticized, because they may be of lower quality[28,29 but see 30-32](McAuley et al 2000, Ferguson & Brannick 2012 but see Cook et al. 1993, Kyzas et al. 2005, Rothstein & Bushman). To reduce this problem, we contacted the owners of large data sets from long-term studies (see also[33] Both et al. 2004). The sample sizes from such data sets allow precise estimates, and are comparable to those from the combined existing literature (see Figures 1c, d). Moreover, study inclusion is not conditional on detection via published results and independent of the phenomenon of interest. In this context, the increased availability of data

due to Open Access practices[34] (Culina et al. 2018) might help unbiased quantification and more objective summaries of existing knowledge.

Evaluating observer bias

Here, we address two types of observer bias that can inflate estimates of assortative mating. First, in studies with multiple observers, pair members may appear more similar if both are measured by the same observer. In our study, this effect was small, but statistically significant. This confounding effect can easily be avoided by limiting observations to a single observer (if feasible), or by calculating correlations between pair members after statistically removing observer effects (see[35] Class et al. 2017 for an elegant solution).

Second, observers may have pre-conceptions about assortative mating, such that measurements suffer from confirmation bias[36] (Nickerson 1998). This is perhaps less likely for data sets that were collected without hypotheses on assortative mating in mind (such as ours). In general, blinding of observers[37] (Holman et al. 2015) is the best countermeasure.

Bias due to temporal and spatial autocorrelation

Our results show that the estimated strength of size-assortative mating is higher when individuals are measured within the same month or at the same site (Figure 1b). For both technical and biological reasons, data may show temporal or spatial autocorrelation and pair members may appear more similar if they are measured closer in time or in space. For example, measures of plumage coloration may show temporal autocorrelation because of changes in the white balance used for calibration of hand-held photo-spectrometers[38] (e.g. Fargevieille et al. 2017) or because plumage color gradually changes after moult due to wear. This can be assessed by quantifying temporal and spatial autocorrelation in measurements, and it can then be controlled for, either statistically or experimentally (e.g. by randomizing measurement order).

Our analyses show that the apparent level of assortative mating may depend on the ecological circumstances in which individuals are measured (Figure 1b). Depending on the research question, this can be of biological interest or it can be a confounding factor. For instance, evolutionary geneticists are interested in assortative mating because it creates gametic phase disequilibrium between loci that affect the trait of interest such as body size[39-42] (Robinson et al. 2017Wilson 1973; Lybch and Walsh 1998; Keller et al. 2013). For this particular purpose it appears most promising to study assortative mating directly at the gene level which directly investigate the assortment of the genetic variants underlying the trait[3,42 but see 43] (Robinson et al.

2017, Domingue et al. 2014 but see Abdellaoui et al. 2014). Another solution may lie in modelling the correlation between pair members in bivariate mixed models[35] (e.g. Class et al. 2017) which could estimate the assortment at different levels by adding different random effects (e.g. disentangle the assortment resulting from time or space, then the correlation of residual is the 'true' assortment).

Assortative mating due to mate choice

To examine whether assortment arises from (mutual) mate choice rather than from other processes, an experimental approach is ideal, in particular when all individuals can be measured before they mate. Where this is not feasible, one could consider a targeted analysis of binary mating decisions observed in the field (e.g. evidence for rejection vs. acceptance of an individual). In the end, knowledge of a study system and the consideration of possible confounds ('like meets like' and 'become alike') may help separating mate choice from confounding factors.

Interpretation of assortative mating in the 9 studied species

How do we interpret the levels of assortment found in our own study species? Is there any evidence that individuals of these species care about the size of their partner? If individuals had a general mating preference for similarly sized partners, we would have expected – given high repeatability of morphological traits –a positive correlation between pair members even when they were measured independently in time or space, but this was not the case (Figure 1b). Assortative mating due to active mate choice could have occurred in some of the species and for some of the traits. Figure S1 shows that Semipalmated sandpipers mated assortatively for wing length ($0.275 < r < 0.295$, $n = 321$, $p < 0.00001$, models 1-3). This might arise because wing length is related to arrival date[44-46](Yong & Moore 1994, 1997, Bowlin 2007) at the arctic breeding grounds in Alaska, and birds pair assortatively by arrival date (likely reflecting 'non-panmixis' rather than active mate choice for similar size;[47] Bearhop et al. 2005). Note that assortative mating was not observed in the same species, studied at another breeding site in Alaska[48](r = -0.08, n = 118, statistical power = 94%; Sandercock 1998, estimate included in 'Web of Science Search').

Figure S1 also shows that Tawny owls mated assortatively for body mass (across three models, r = 0.17, 0.25 and 0.34, respectively, n = 351, Table S2 to S4). This could result from the mechanism of 'becoming alike'[10,12,49] (Price and Vandenberg 1980; Burleson and Denton 1992; Anderson et al. 2003). Pair members were typically weighed about one month after egg-laying,

during a period when incubating females lose weight[50] (Karell et al. 2011) and are provided with food by their partner[51] (Brommer et al. 2015). Variation in male hunting success might hence explain the similarity of partners when the male and female were measured close in time ('nearest' model), and the lower values of assortative mating (r = 0.17) in the 'random alignment model' (where comparison is often between years; Figure S1).

Assortative mating for traits other than size

Our meta-analysis (Figure 1) shows that the highest levels of assortative mating are for age (r = 0.41), behavioral traits (r = 0.33), physiological traits (r = 0.30), and plumage traits (r = 0.26). Assortment by age might be most parsimoniously explained by a lack of panmixis ('like meets like')[52] (Ferrer & Penteriani 2003). For instance, in species with long-term pair-bonds, new pairs are typically formed among first-time breeders, not necessarily because of active choice but as a consequence of probability of encounter between unpaired individuals. For behavioral and physiological traits, which are more flexible than morphological traits, the correlation is presumably more strongly affected by shared environmental effects (time and space) and by the partners influencing each other ('become alike')[10,35,53-55] (Anderson et al. 2003, Class et al. 2017, Duckworth and Kruuk 2009, Gimelfarb 1988a, b). In highly flexible traits such as behavioral phenotypes, one can expect the largest confounding effects of environment and measurement error[35,56,57] (Dingemanse et al. 2012, Bell et al. 2009, Class et al. 2017).Finally, assortative mating for plumage traits might be caused by mate choice for similar phenotypes. Indeed, some bird species that have evolved a striking polymorphism in plumage coloration show a clear pattern of assortative mating for color type mediated by sexual imprinting on parental phenotypes[58-61] (Cooch 1959, Odonald 1959, Findlay et al 1985, Bonneaud et al. 2006). Because plumage coloration facilitates species recognition, it appears plausible that assortative mating by color morph results from an imprinting mechanism that has evolved to prevent heterospecific mating in general.

The evolution of mate choice for similarity

Assortative mating can arise from mate choice by two different processes. First, individuals may mate assortatively by indicators of phenotypic quality, because high-quality individuals would only accept a high-quality partner. However, in socially monogamous species, such as the majority of birds, selection may not favour strong choosiness, because the costs of competing for a more ornamented partner could exceed the benefits from such choosiness, especially if ornaments are not highly reliable indicators of receivable fitness gains[21] (Wang et al. 2017). Second, assortment could arise from mate choice for phenotypic similarity. This would result in

lower levels of competition for mates, because preferences diverge between individuals. However, a tendency to mate assortatively across many dimensions of phenotypic variance that exists within species (i.e. effectively in all traits that have been quantified; see Figure 1a) would often result in close inbreeding, because relatedness leads to similarity. Hence, such preferences might be selected against because inbreeding is usually detrimental[62,63](Keller et al. 1998, Keller & Waller 2002).

In contrast to the above scenarios, some traits like personality characteristics could be important for behavioral compatibility of the pair, which could in turn lead to better parental care and hence higher reproductive success. The benefits of compatibility might thus outweigh an increased risk of inbreeding plus search costs for finding a compatible partner [20,64] (Figueredo et al. 2006, Ihle et al. 2015). In this context, the mechanism of convergence between pair members ('become alike') may also deserve more attention. Convergence in behavioural phenotypes could serve an adaptive function if it reduces conflict among pair members and increases pair bond stability[65,66] (Acitelli et al. 2001; Gonzaga et al. 2007).

Conclusions

Assortative mating for certain phenotypic traits is an interesting biological phenomenon that deserves attention. However, our results show that it is not necessarily an outcome of mate choice, as is sometimes implied, and it might be less strong than meta-analyses of the published literature suggest. We argue for careful consideration of alternative mechanisms and confounding effects. Doing this may lead to the conclusion that the pattern of assortative mating was 'spurious', but it may also lead to deeper insight. Finally, our study suggests that greater use could be made of large published or unpublished datasets from long-term studies: incorporating such data into meta-analysis might lead to more trustworthy conclusions.

Methods

Published data

Literature search and inclusion criteria - In March 2015 we searched for published literature on assortative mating in birds using Web of Science with the key-words "birds" and "*assortative mating" (which also covers the term "disassortative mating"). This resulted in 406 hits, of which 129 studies focused on assortative mating within populations (as opposed to studies on the mixing of two defined populations, e.g. in hybrid zones). The 129 studies contained 536

estimates of the strength of assortative mating for any phenotype from 106 species. We refer to these data as 'Web of Science Search'.

Next, we identified additional studies (missed in the 'Web of Science Search') by manually screening the introduction and discussion sections of the 129 publications mentioned above. Of the 66 additional studies identified, 29 concerned assortative mating within populations. These 29 studies contained 88 estimates of assortative mating from 27 species. We refer to this dataset as 'Cited Studies'.

Data extraction and categorization - From both datasets, we extracted the Pearson correlation coefficient (r) as an estimate of the strength of assortative mating between pair members. If 'r' was missing, we calculated it from the following three test statistics.

F-test with a single numerator degree of freedom and denominator degrees of freedom (df):

$$r = \sqrt{\frac{F(1,-)}{F(1,-)+df}}$$ [67](Coltman & Slate 2003)

χ^2 statistic with one degree of freedom and sample size (n):

$$r = \sqrt{\frac{\chi 2}{n}}$$ [67](Coltman & Slate 2003)

For studies in which the strength of assortment was reported in 2 x 2 contingency tables (e.g. two different plumage types), we calculated r following Nakagawa and Cuthill[68] (2007, Table 2, Equation 9):

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

where A, B, C, and D represent the observed cell frequencies, and n = A + B + C + D = the total sample size.

To avoid pseudoreplication we checked multiple studies on the same species (especially those from the same research group), and excluded redundant estimates from the same population and same period, giving priority to the estimate based on the largest sample size.

We classified the phenotypic traits for which assortative mating had been reported into one of eight trait categories: body size (n = 357 estimates), body condition (n = 23), plumage coloration

(n = 145), age (n = 38), behaviour (n = 32), physiology (n = 9), heterozygosity (n = 9), and other (n = 11). Here, we focus on the best-documented assortment by body size traits.

Estimating assortative mating - To estimate assortative mating, Pearson's r - weighed by sample size ($(n-3)^{0.5}$, where n is the number of pairs[68], Nakagawa & Cuthill 2007) - was modelled as the dependent variable, with 'Type of trait' as a fixed effect (factor with eight levels), and 'Species' and 'Study' (i.e. publication) as random effects. We removed the intercept to obtain parameter estimates and 95% confidence intervals (95% CI) for each trait.

Unpublished data from long-term field studies

Selection of studies **-** To obtain data from comparable field studies, that have not gone through the filtering steps of publication and detection via search terms or citation, we contacted 10 researchers who run long-term field studies. All but one agreed to provide the raw data, yielding nine data sets from 8 different species and from 4-38 years of study (see Supplementary Methods for details). The studies were chosen based on personal contacts, independent of knowledge about mate choice, but with the aim to include both non-passerines (n = 3) and passerines (n = 5; Table 1).

Given these selection criteria, we expect no bias with regard to assortative mating. All data sets were analysed using the same predefined methods. Our aim was to use these data in two ways: (1) to compare with data from the literature search (see above) to assess the extent of search and reporting bias, and (2) to quantify the extent to which correlations among pair members are affected by shared confounding effects (observer bias, temporal and spatial autocorrelation). For the latter analyses, not all data sets contained all necessary information, but we used all available information irrespective of the outcome of the analysis.

Data handling **-** The unpublished data consist of two tables. (1) Supplementary file 1 lists all the pairs that have been identified across the nine studies where both pair members have at least one morphological record (n = 6,309, including repeated records from different years). This dataset also includes latitude and longitude of the nest site (Lambert azimuthal equal-area projection, units = meters) and year and, if available, the putative date of the first egg. (2) Supplementary file 2 lists all available records of morphological traits (n = 41,896 which covered more than 95% of individuals included in (1), see Table 1) .This dataset also includes the location where the individual was caught, the date of catching, and the observer who measured the individual.

We then combined the information from data tables (1) and (2) in Supplementary File 3. We selected the first record (closest to pair formation) of all unique pairs (n = 5,199; Table 1). In most pairs (65.2%, see Table 1) one or both partners had been measured repeatedly for a given trait (regardless of whether they were paired at the time of measurement). For example, the female might have been weighed twice and the male three times. In this case, there are six combinations to align the measurements of the partners (2 x 3). The number of such combinations per pair (range 1 to 196) varied between studies (mean 4.4, median = 2, Table 1) and allowed for a total of 72,739 combinations of male measurement by homologous female measurement.

Each of these combinations can be characterized by the circumstances of measurement (place, time, and observer) for each of the partners. We considered the pair members as measured at the 'same site' (n = 26,542, 36.5%) if the Euclidean distance between their sites of capture (usually the nest) was less than 10 m, or at 'different sites' if they were caught more than 10 m apart (n = 44,112, 60.6%; the remaining 2.9% were cases of missing information). Likewise, combinations of measurements were defined as from the 'same month' (n = 20,729, 28.5%) if obtained less than 30 days apart or as from 'different months' (n = 51,995, 71.5%; n = 15 cases of missing data). Combinations of measurements were either from the 'same observer' (n = 35,018, 48.1%) or from 'different observers' (n = 24,771, 34.1%; 17.8% are missing data). For each of the 16,543 unique pair-trait combinations we also selected the combination of measurements from the pairs' first year of breeding.

Estimating assortative mating **-** We estimated the strength of assortative mating by calculating Pearson correlation coefficients (r) and their 95% CI using 6 different approaches (models 1-6 below) that essentially differ in how the available morphological measurements are used.

The 'random alignment model' (model 1)

For each combination of study and trait (n = 32), we first randomly sampled (1,000 times) from each pair one of the available male-female combinations of measurements and then calculated r and its 95% CI (averaged across the 1,000 replicates). We then summarized the 32 average correlation coefficients, weighed by sample size $(n-3)^{0.5}$, where n is the number of pairs, using a mixed effect model with 'Study' and 'Trait' as random effects. This approach reflects the strength of assortment under 'random' measuring conditions to the extent allowed by the data, i.e. given that 38% of the data were still from the same site, 29% from the same month, and 59% from the same observer.

The 'average model' (model 2)

Similar to model 1, but calculating r-values (and 95% CI) using mean trait values for each individual. This approach of averaging all available measurements approximates the individuals' average phenotype (approach similar to the one used in quantitative genetics to estimate the underlying breeding value).

The 'nearest model' (model 3)

Similar to model 1, but using the measurements taken closest in time to pair formation (see above) to calculate r-values (and their 95 % CI) between pair members. This approach reflects the phenotypes around the time of pair formation, when mate choice can take place.

Model to reveal observer effect (model 4)

For each study-trait combination where multiple observers had contributed data (n = 22 out of the 32 study-trait combinations, excluding barn swallows and western bluebirds), we calculated two r-values: one that included all pair combinations measured by the 'same observer' (22 correlations, $n_{combinations}$: range = 161-5,837, mean = 1,514) and one across all pair combinations measured by 'different observers' (22 correlations, $n_{combinations}$: range = 70-3,227, mean = 1,172). These 44 correlation coefficients were summarized in a mixed model as described above (weighted by the number of pair combinations): 'Study', 'Trait' and 'Study-trait combination' were added as random effects with 'observer category' (same or different) as the fixed effect of interest.

Model to reveal temporal autocorrelation effect (model 5)

Similar to model 4, but contrasting r-values from pairs where the members had been measured in the 'same month' (32 correlations, $n_{combinations}$: range = 58-3,064, mean = 648) versus in 'different months' (31 correlations, $n_{combinations}$: range = 223-6,204, mean = 1,677).

Model to reveal spatial autocorrelation effects (model 6)

Similar to model 4, but contrasting r-values from pairs where the members had been measured at the 'same site' (32 correlations, $n_{combinations}$: range = 210-2,773, mean = 829) versus at 'different sites' (32 correlations, $n_{combinations}$: range = 7-6,201, mean = 1,379).

Experimental data on zebra finches

We assessed assortative mating for size using captive populations of zebra finches (*Taeniopygia guttata*). As a rule, in each experiment, all birds were measured by a single observer prior to their release into breeding aviaries. Measurements were taken in an order that was independent of allocation to aviaries. This excludes systematic observer error as well as spatial and temporal heterogeneity. To minimize the 'scale-of-choice-effect'[8,83] (Rolan-Alvarez et al. 2015; Ng et al. 2016), we analysed the degree of assortative mating within aviaries, hence comprising only the birds that were available for pairing at the time of release. To avoid selective reporting, we summarize all available information from our laboratory (partly published in[21] Wang et al. 2017), comprising five experiments that largely fulfil the above.

References

1 Alford, J. R., Hatemi, P. K., Hibbing, J. R., Martin, N. G. & Eaves, L. J. The politics of mate choice. *The Journal of Politics* **73**, 362-379 (2011).

2 Klofstad, C. A., McDermott, R. & Hatemi, P. K. The dating preferences of liberals and conservatives. *Political Behavior* **35**, 519-538 (2013).

3 Domingue, B. W., Fletcher, J., Conley, D. & Boardman, J. D. Genetic and educational assortative mating among US adults. *P Natl Acad Sci USA* **111**, 7996-8000, doi:10.1073/pnas.1321426111 (2014).

4 Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nat Hum Behav* **1**, doi:UNSP 001610.1038/s41562-016-0016 (2017).

5 Tenesa, A., Rawlik, K., Navarro, P. & Canela-Xandri, O. Genetic determination of height-mediated mate choice (vol 16, 269, 2016). *Genome Biol* **17**, doi:ARTN 2610.1186/s13059-016-0894-3 (2016).

6 Stulp, G., Simons, M. J. P., Grasman, S. & Pollet, T. V. Assortative mating for human height: A meta-analysis. *Am J Hum Biol* **29**, doi:ARTN e2291710.1002/ajhb.22917 (2017).

7 Jiang, Y. X., Bolnick, D. I. & Kirkpatrick, M. Assortative Mating in Animals. *Am Nat* **181**, E125-E138, doi:10.1086/670160 (2013).

8 Rolan-Alvarez, E. *et al.* The scale-of-choice effect and how estimates of assortative mating in the wild can be biased due to heterogeneous samples. *Evolution* **69**, 1845-1857, doi:10.1111/evo.12691 (2015).

9 Village, A. Spring arrival times and assortative mating of kestrels in south Scotland. *The Journal of Animal Ecology*, 857-868 (1985).

10 Anderson, C., Keltner, D. & John, O. P. Emotional convergence between people over time. *J Pers Soc Psychol* **84**, 1054-1068, doi:10.1037/0022-3514.84.5.1054 (2003).

11 Feunekes, G. I. J., Stafleu, A., de Graaf, C. & van Staveren, W. A. Family resemblance in fat intake in the Netherlands. *Eur J Clin Nutr* **51**, 793-799, doi:DOI 10.1038/sj.ejcn.1600494 (1997).

12 Price, R. A. & Vandenberg, S. G. Spouse Similarity in American and Swedish Couples. *Behav Genet* **10**, 59-71, doi:Doi 10.1007/Bf01067319 (1980).

13 García-Navas, V., Ortego, J. & Sanz, J. J. Heterozygosity-based assortative mating in blue tits (Cyanistes caeruleus): implications for the evolution of mate choice. *Proceedings of the Royal Society of London B: Biological Sciences* **276**, 2931-2940 (2009).

14 Houtman, A. M. & Falls, J. B. Negative assortative mating in the white-throated sparrow, Zonotrichia albicollis: the role of mate choice and intra-sexual competition. *Animal Behaviour* **48**, 377-383 (1994).

15 Cunningham, R., Lindenmayer, D., Nix, H. & Lindenmayer, B. Quantifying observer heterogeneity in bird counts. *Austral Ecology* **24**, 270-277 (1999).

16 Open-Science-Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

17 Greenwald, A. G. Consequences of Prejudice against Null Hypothesis. *Psychol Bull* **82**, 1-19, doi:DOI 10.1037/h0076157 (1975).

18 Kulshrestha, J. *et al.* in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* 417-432 (ACM).

19 Ferguson, C. J. & Heene, M. A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspect Psychol Sci* **7**, 555-561 (2012).

20 Ihle, M., Kempenaers, B. & Forstmeier, W. Fitness Benefits of Mate Choice for Compatibility in a Socially Monogamous Species. *Plos Biol* **13**, doi:ARTN e100224810.1371/journal.pbio.1002248 (2015).

21 Wang, D. P., Forstmeier, W. & Kempenaers, B. No mutual mate choice for quality in zebra finches: Time to question a widely held assumption. *Evolution* **71**, 2661-2676, doi:10.1111/evo.13341 (2017).

22 Tang, J. L. & Liu, J. L. Y. Misleading funnel plot for detection of bias in meta-analysis. *J Clin Epidemiol* **53**, 477-484, doi:Doi 10.1016/S0895-4356(99)00204-8 (2000).

23 Fanelli, D. Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *Plos One* **5**, doi:ARTN e1027110.1371/journal.pone.0010271 (2010).

24 Kittelman, A., Gion, C., Horner, R. H., Levin, J. R. & Kratochwill, T. R. Establishing Journalistic Standards for the Publication of Negative Results. *Remedial and Special Education*, 0741932517745491 (2018).

25 Coltman, D. & Slate, J. Microsatellite measures of inbreeding: a meta-analysis. *Evolution* **57**, 971-983 (2003).

26 Wang, D. *et al.* Irreproducible text-book 'knowledge': the effects of color bands on zebra finch fitness *Evolution (accepted)*, doi:DOI: https://doi.org/10.5061/dryad.cc145b6 (2018).

27 Sanchez-Tojar, A. *et al.* Meta-analysis challenges a textbook example of status signalling: evidence for publication bias. *bioRxiv*, 283150 (2018).

28 McAuley, L., Pham, B., Tugwell, P. & Moher, D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* **356**, 1228-1231, doi:Doi 10.1016/S0140-6736(00)02786-0 (2000).

29 Ferguson, C. J. & Brannick, M. T. Publication Bias in Psychological Science: Prevalence, Methods for Identifying and Controlling, and Implications for the Use of Meta-Analyses. *Psychol Methods* **17**, 120-128, doi:10.1037/a0024445 (2012).

30 Cook, D. J. *et al.* Should Unpublished Data Be Included in Metaanalyses - Current Convictions and Controversies. *Jama-J Am Med Assoc* **269**, 2749-2753, doi:DOI 10.1001/jama.269.21.2749 (1993).

31 Kyzas, P. A., Loizou, K. T. & Ioannidis, J. P. A. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer I* **97**, 1043-1055, doi:10.1093/jnci/dji184 (2005).

32 Rothstein, H. R. & Bushman, B. J. Publication Bias in Psychological Science: Comment on Ferguson and Brannick (2012). *Psychol Methods* **17**, 129-136, doi:10.1037/a0027128 (2012).

33 Both, C. *et al.* Large-scale geographical variation confirms that climate change causes birds to lay earlier. *P R Soc B* **271**, 1657-1662, doi:10.1098/rspb.2004.2770 (2004).

34 Culina, A. *et al.* Navigating the unfolding open data landscape in ecology and evolution. *Nature ecology & evolution* **2**, 420 (2018).

35 Class, B., Dingemanse, N. J., Araya-Ajoy, Y. G. & Brommer, J. E. A statistical methodology for estimating assortative mating for phenotypic traits that are labile or measured with error. *Methods Ecol Evol* **8**, 1910-1919, doi:10.1111/2041-210x.12837 (2017).

36 Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* **2**, 175 (1998).

37 Holman, L., Head, M. L., Lanfear, R. & Jennions, M. D. Evidence of experimental bias in the life sciences: why we need blind data recording. *Plos Biol* **13**, e1002190 (2015).

38 Fargevieille, A., Grégoire, A., Charmantier, A., Rey Granado, M. & Doutrelant, C. Assortative mating by colored ornaments in blue tits: space and time matter. *Ecology and evolution* **7**, 2069-2078 (2017).

39 Wilson, S. The correlation between relatives under the multifactorial model with assortative mating. *Annals of human genetics* **37**, 189-204 (1973).

40 Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. Vol. 1 (Sinauer Sunderland, MA, 1998).

41 Keller, M. C. *et al.* The genetic correlation between height and IQ: Shared genes or assortative mating? *PLoS genetics* **9**, e1003451 (2013).

42 Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**, 0016 (2017).

43 Abdellaoui, A., Verweij, K. J. H. & Zietsch, B. P. No evidence for genetic assortative mating beyond that due to population stratification. *P Natl Acad Sci USA* **111**, E4137-E4137, doi:10.1073/pnas.1410781111 (2014).

44 Yong, W. & Moore, F. R. Flight Morphology, Energetic Condition, and the Stopover Biology of Migrating Thrushes. *Auk* **111**, 683-692 (1994).

45 Yong, W. & Moore, F. R. Spring stopover of intercontinental migratory thrushes along the northern coast of the Gulf of Mexico. *Auk* **114**, 263-278, doi:Doi 10.2307/4089167 (1997).

46 Bowlin, M. S. Sex, wingtip shape, and wing-loading predict arrival date at a stopover site in the Swainson's thrush (Catharus ustulatus). *The Auk* **124**, 1388-1396 (2007).

47 Bearhop, S. *et al.* Assortative mating as a mechanism for rapid evolution of a migratory divide. *Science* **310**, 502-504, doi:10.1126/science.1115661 (2005).

48 Sandercock, B. K. Assortative mating and sexual size dimorphism in western and semipalmated sandpipers. *Auk* **115**, 786-791, doi:Doi 10.2307/4089430 (1998).

49 Burleson, B. R. & Denton, W. H. A New Look at Similarity and Attraction in Marriage - Similarities in Social-Cognitive and Communication-Skills as Predictors of Attraction and Satisfaction. *Commun Monogr* **59**, 268-287, doi:Doi 10.1080/03637759209376269 (1992).

50 Karell, P. *et al.* Blood parasites mediate morph-specific maintenance costs in a colour polymorphic wild bird. *J Evolution Biol* **24**, 1783-1792, doi:10.1111/j.1420-9101.2011.02308.x (2011).

51 Brommer, J. E., Karell, P., Aaltonen, E., Ahola, K. & Karstinen, T. Dissecting direct and indirect parental effects on reproduction in a wild bird of prey: dad affects when but not how much. *Behav Ecol Sociobiol* **69**, 293-302, doi:10.1007/s00265-014-1842-4 (2015).

52 Ferrer, M. & Penteriani, V. A process of pair formation leading to assortative mating: passive age-assortative mating by habitat heterogeneity. *Animal Behaviour* **66**, 137-143, doi:10.1006/anbe.2003.2158 (2003).

53 Duckworth, R. A. & Kruuk, L. E. B. Evolution of Genetic Integration between Dispersal and Colonization Ability in a Bird. *Evolution* **63**, 968-977, doi:10.1111/j.1558-5646.2009.00625.x (2009).

54 Gimelfarb, A. Processes of Pair Formation Leading to Assortative Mating in Biological Populations - Encounter-Mating Model. *Am Nat* **131**, 865-884, doi:Doi 10.1086/284827 (1988).

55 Gimelfarb, A. Processes of Pair Formation Leading to Assortative Mating in Biological Populations - Dynamic Interaction-Model. *Theor Popul Biol* **34**, 1-23, doi:Doi 10.1016/0040-5809(88)90032-9 (1988).

56 Dingemanse, N. J. *et al.* Variation in personality and behavioural plasticity across four populations of the great tit Parus major. *J Anim Ecol* **81**, 116-126, doi:10.1111/j.1365-2656.2011.01877.x (2012).

57 Bell, A. M., Hankison, S. J. & Laskowski, K. L. The repeatability of behaviour: a meta-analysis. *Animal Behaviour* **77**, 771-783, doi:10.1016/j.anbehav.2008.12.022 (2009).

58 Cooch, F. G. & Beardmore, J. A. Assortative Mating and Reciprocal Difference in the Blue-Snow Goose Complex. *Nature* **183**, 1833-1834, doi:DOI 10.1038/1831833b0 (1959).

59 Odonald, P. Possibility of Assortive Mating in the Arctic Skua. *Nature* **183**, 1210-1211, doi:DOI 10.1038/1831210a0 (1959).
60 Findlay, C. S., Rockwell, R. F., Smith, J. A. & Cooke, F. Life-History Studies of the Lesser Snow Goose (Anser-Caerulescens-Caerulescens) .6. Plumage Polymorphism, Assortative Mating and Fitness. *Evolution* **39**, 904-914, doi:Doi 10.2307/2408689 (1985).
61 Bonneaud, C., Chastel, O., Federici, P., Westerdahl, H. & Sorci, G. Complex Mhc-based mate choice in a wild passerine. *P R Soc B* **273**, 1111-1116, doi:10.1098/rspb.2005.3325 (2006).
62 Keller, L. F. Inbreeding and its fitness effects in an insular population of song sparrows (Melospiza melodia). *Evolution* **52**, 240-250 (1998).
63 Keller, L. F. & Waller, D. M. Inbreeding effects in wild populations. *Trends Ecol Evol* **17**, 230-241 (2002).
64 Figueredo, A. J., Sefcek, J. A. & Jones, D. N. The ideal romantic partner personality. *Pers Indiv Differ* **41**, 431-441, doi:10.1016/j.paid.2006.02.004 (2006).
65 Gonzaga, G. C., Campos, B. & Bradbury, T. Similarity, convergence, and relationship satisfaction in dating and married couples. *J Pers Soc Psychol* **93**, 34-48, doi:10.1037/0022-3514.93.1.34 (2007).
66 Acitelli, L. K., Kenny, D. A. & Weiner, D. The importance of similarity and understanding of partners' marital ideals to relationship satisfaction. *Pers Relationship* **8**, 167-185, doi:DOI 10.1111/j.1475-6811.2001.tb00034.x (2001).
67 Coltman, D. W. & Slate, J. Microsatellite measures of inbreeding: A meta-analysis. *Evolution* **57**, 971-983 (2003).
68 Nakagawa, S. & Cuthill, I. C. 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 (2007).
69 Petrzelkova, A. *et al.* Brood parasitism and quasi-parasitism in the European barn swallow Hirundo rustica rustica. *Behav Ecol Sociobiol* **69**, 1405-1414, doi:10.1007/s00265-015-1953-6 (2015).
70 Wilkins, M. R. *et al.* Phenotypic differentiation is associated with divergent sexual selection among closely related barn swallow populations. *J Evolution Biol* **29**, 2410-2421, doi:10.1111/jeb.12965 (2016).
71 Drummond, H., Torres, R. & Krishnan, V. V. Buffered development: Resilience after aggressive subordination in infancy. *Am Nat* **161**, 794-807, doi:Doi 10.1086/375170 (2003).
72 Kiere, L. M., Ramos, A. G. & Drummond, H. No evidence that genetic compatibility drives extra-pair behavior in female blue-footed boobies. *J Avian Biol* **47**, 871-879, doi:10.1111/jav.01061 (2016).
73 Delhey, K., Johnsen, A., Peters, A., Andersson, S. & Kempenaers, B. Paternity analysis reveals opposing selection pressures on crown coloration in the blue tit (Parus caeruleus). *P R Soc B* **270**, 2057-2063, doi:10.1098/rspb.2003.2460 (2003).
74 Foerster, K., Delhey, K., Johnsen, A., Lifjeld, J. T. & Kempenaers, B. Females increase offspring heterozygosity and fitness through extra-pair matings. *Nature* **425**, 714-717, doi:10.1038/nature01969 (2003).
75 Schlicht, L., Girg, A., Loes, P., Valcu, M. & Kempenaers, B. Male extrapair nestlings fledge first. *Animal Behaviour* **83**, 1335-1343, doi:10.1016/j.anbehav.2012.02.021 (2012).
76 Araya-Ajoy, Y. G. *et al.* Sources of (co)variation in alternative siring routes available to male great tits (Parus major). *Evolution* **70**, 2308-2321, doi:10.1111/evo.13024 (2016).
77 Both, C. *et al.* Delayed age at first breeding and experimental removals show large non-breeding surplus in Pied Flycatchers. *Ardea* **105**, 43-60, doi:10.5253/arde.v105i1.a2 (2017).

78 Bulla, M., Valcu, M., Rutten, A. L. & Kempenaers, B. Biparental incubation patterns in a high-Arctic breeding shorebird: how do pairs divide their duties? *Behav Ecol* **25**, 152-164, doi:10.1093/beheco/art098 (2014).

79 Bulla, M. *et al.* Flexible parental care: Uniparental incubation in biparentally incubating shorebirds. *Sci Rep-Uk* **7**, doi:ARTN 1285110.1038/s41598-017-13005-y (2017).

80 Karell, P., Ahola, K., Karstinen, T., Zolei, A. & Brommer, J. E. Population dynamics in a cyclic environment: consequences of cyclic food abundance on tawny owl reproduction and survival. *J Anim Ecol* **78**, 1050-1062, doi:10.1111/j.1365-2656.2009.01563.x (2009).

81 Aguillon, S. M. & Duckworth, R. A. Kin aggression and resource availability influence phenotype-dependent dispersal in a passerine bird. *Behav Ecol Sociobiol* **69**, 625-633, doi:10.1007/s00265-015-1873-5 (2015).

82 Duckworth, R. A. & Semenov, G. A. Hybridization Associated with Cycles of Ecological Succession in a Passerine Bird. *Am Nat* **190**, E94-E105, doi:10.1086/693160 (2017).

83 Ng, T. P., Williams, G. A., Davies, M. S., Stafford, R. & Rolán-Alvarez, E. Sampling scale can cause bias in positive assortative mating estimates: evidence from two intertidal snails. *Biological journal of the Linnean Society* **119**, 414-419 (2016).

Funding

This work was supported by the Max Planck Society (to B.K.) and the China Scholarship Council (CSC; stipend to D.W.). Work for the monitoring of Blue-footed boobies on Isla Isabel was supported by the Universidad Nacional Autónoma de México (PAPIIT, IN211491, IN-200702-3, IN206610-3, IN205313), the Consejo Nacional de Cienciay Tecnología (CONACYT, 81823, 47599, 34500-V, 4722-N9407, D112-903581, 31973H and 104313) and the National Geographic Society to Dr. Hugh Drummond. Work of Western bluebirds was supported by US National Science Foundation grants (DEB-0918095 and DEB-1350107 to R.A.D.). Work of Pied flycatcher was supported by a VIDI-grant from the Netherlands Organisation for Scientific Research (NWO) to C.B. Work of Tawny owl was supported by the Academy of Finland (project 314108) to P.K. Work of Barn swallow was supported by the Czech Science Foundation project No. 15-11782S to T.A.

Acknowledgements

We thank Hugh Drummond and Jon Brommer for help with data access. For the data on Barn swallows, we thank Adela Petrzelkova, Oldrich Tomasek, Romana Michalkova, Marie Adamkova and Jana Abrechtova for their contribution. We also thank local farm owners for providing access to breeding colonies, specifically the Kotrba family at Hamr farm, the Kraus family at Šaloun Farm and the Pulec family. For the data on Blue-footed boobies, we thank Ale Ramos for data collection. We also thank Alejandra G. Ramos who was instrumental in taking the measurements used in the analyses included in this paper. We are grateful to the many dozens of volunteers who participated both in taking the body measurements used in this paper and in the yearly monitoring for the long-term database, as well as to the fishermen of Isla Isabel for their friendship and assistance. For the data on Blue tits and Semipalmated sandpipers, we thank all the people from the Max Planck Institute for Ornithology who collected data in the field. For the data on Great tits, we thank all people from the Behavioural Ecology group of the Department of Biology, Ludwig Maximilians University of Munich who contributed to data collection. For the data on Pied flycatchers, we thank Richard Ubels who is responsible for managing the data base, and Rob Bijlsma who did many measurements. For the data on Tawny owls, we thank Kari Ahola and Teuvo Karstinen for their inexhaustible data collection efforts and willingness to share these data. We are also grateful for the efforts of all other members of Kimpari Bird Projects (KBP) in collecting data on tawny owls during all these years. For the data on Western bluebirds, we thank residents of the Hayes Creek neighborhood for kindly allowing us to monitor nests on their properties. For the data on Zebra finches, we thank Katrin Martin,

Malika Ihle, Sanja Janker, Johannes Schreiber, and Ulrich Knief for data collection and all animal care takers for their help.

Author contributions

D.W. and W.F. conceived the project with input of B.K.. W.F., M.V. and D.W. analysed the data. D.W. and W.F. coordinated the project with input from N.D.. Data was contributed by T.A. (Barn swallow), L.M.K. (Blue-footed booby), B.K. and M.V. (Blue tit), N.D. (Great tit), C.B. (Pied flycatcher), B.K. and M.B. (Semipalmated sandpiper), P.K. (Tawny owl), R.A.D. (Western bluebirds), B.K., W.F. and D.W. (Zebra finch). D.W. and W.F wrote the first draft. All co-authors contributed to writing the manuscript. D.W., W.F, and M.B. integrated the comments from the other co-authors. B.K., W.F., M.B. and D.W. wrote the final draft.

Table 1. Overview of the 'Unpublished data' from nine long-term field studies. For each population we give its abbreviation (Abbr.), the country where the study site is located, a reference for more details about the study, the duration of the study, the number of unique pairs where both members were measured at least once, the proportion of pairs for which multiple morphological measurements were available for at least one member, the average number of male-measurement by female-measurement combinations that can be created per pair (e.g. male partner measured 2 times, female partner measured 3 times leads to 2×3= 6 combinations), and the availability (indicated with Y) of morphological data (C = culmen length, M = body mass, U = ulna length, L = tail length, T = tarsus length, W = wing length, P = length of primary 3, H = length of head including culmen). Overall, data include 32 population-trait combinations and 16,543 pair-trait combinations from a total of 5,199 pairs.

Species name	Abbr.	Country	Ref.	Years	n unique pairs	% multiple measurements	n combi-nations	C	M	U	L	T	W	P	H
Barn swallow *Hirundo rustica*	BS	Czech Republic	[69, 70]	6	235	63.0%	2.7		Y		Y	Y	Y		
Blue-footed booby *Sula nebouxii*	BB	Mexico	[71, 72]	4	510	20.5%	1.4	Y	Y	Y					
Blue tit *Cyanistes caeruleus*	BT_K	Austria	[73, 74]	9	332	90.6%	11.8		Y			Y	Y		
Blue tit *Cyanistes caeruleus*	BT_W	Germany	[75]	7	511	81.5%	5.5		Y			Y		Y	
Great tit *Parus major*	GT	Germany	[76]	6	814	66.0%	3.4		Y			Y		Y	
Pied flycatcher *Ficedula hypoleuca*	PF	Holland	[77]	9	1832	76.7%	4.1		Y			Y		Y	
Semipalmated sandpiper *Calidris pusilla*	SS	USA	[78, 79]	7	325	49.8%	2.0	Y	Y			Y	Y		Y
Tawny owl *Strix aluco*	TO	Finland	[80]	38	350	83.3%	11.6		Y		Y		Y		
Western bluebird *Sialia mexicana*	WB	USA	[81, 82]	15	290	55.4%	2.1	Y	Y		Y	Y	Y		
Total					5199	65.2%									

Figure 1. (a) The magnitude of assortative mating in birds for various types of traits based on a meta-analysis of the published literature. Shown are Pearson's correlation coefficients (r). Dots represent mean values, bars the 95% CI (based on Table S1). 'n' indicates the number of estimates for a given trait category followed by the number of pair-trait combinations in parentheses. The data comprises both the 'Web of Science Search' and the 'Cited studies' (see methods). The dotted line indicates no assortative mating (r = 0), negative r-values indicate disassortative mating, and positive r-values indicate assortative mating. (b) Strength of assortative mating for size as a function of data source. Shown are mean Pearson's r and 95% CI. Sample sizes are indicated as in (a). Searching the 'Web of Science' for keywords yielded a weighted mean estimate of assortative mating for size of r = 0.201 ± 0.022 (referred to as 'Web of Science Search'; t = 9.02, p < 0.0001). Published studies that had been missed by the Web of Science search, but were detected because they had been cited by the former set of studies (referred to as 'Cited studies'), yielded a somewhat lower estimate of r = 0.135 ± 0.043 (t = 3.11, p = 0.002). The weighted mean estimate from our unpublished field data was even lower (referred to as 'Unpublished data', r = 0.106 ± 0.049, t = 2.17, p = 0.03) when using the 'nearest model' (the most frequently used method in the published literature). Finally, the 'Experimental data' on zebra finches in which we controlled for confounding factors, suggests the absence of assortative mating r = -0.003 ± 0.141 (t = -0.02, p = 1.0, Table S7). (c) Strength of assortative mating calculated from 'Unpublished data' (nine long-term field studies) as a function of measurement context. Assortative mating (Pearson r ± 95% CI) is stronger when the measurements of the two partners were taken by the same observer, within the same month, or at the same site, compared to measures taken by different observers, in different months (>30 days apart), or at different sites (>10 m apart) (Table S5). Sample sizes are indicated as in (a).

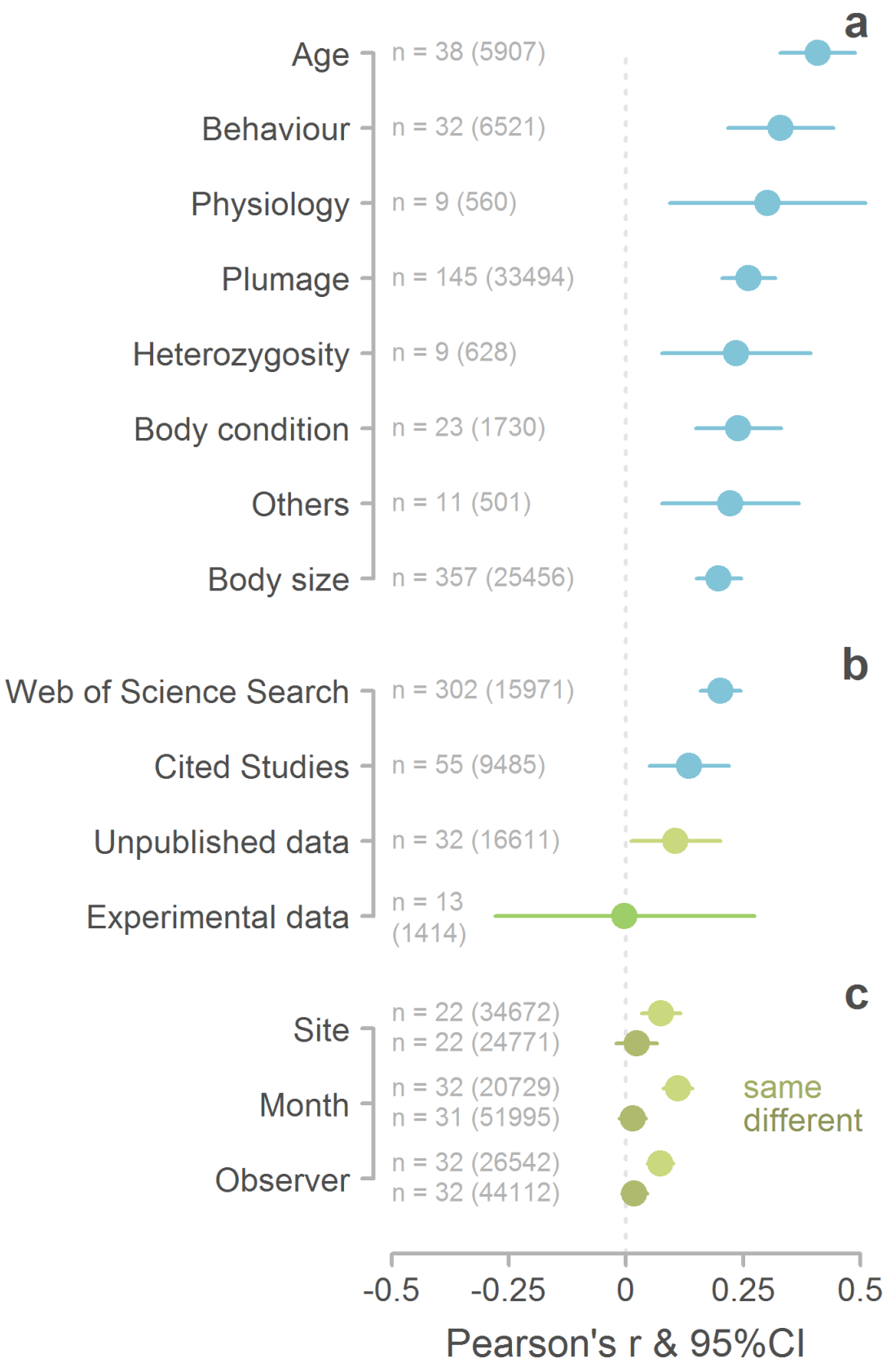
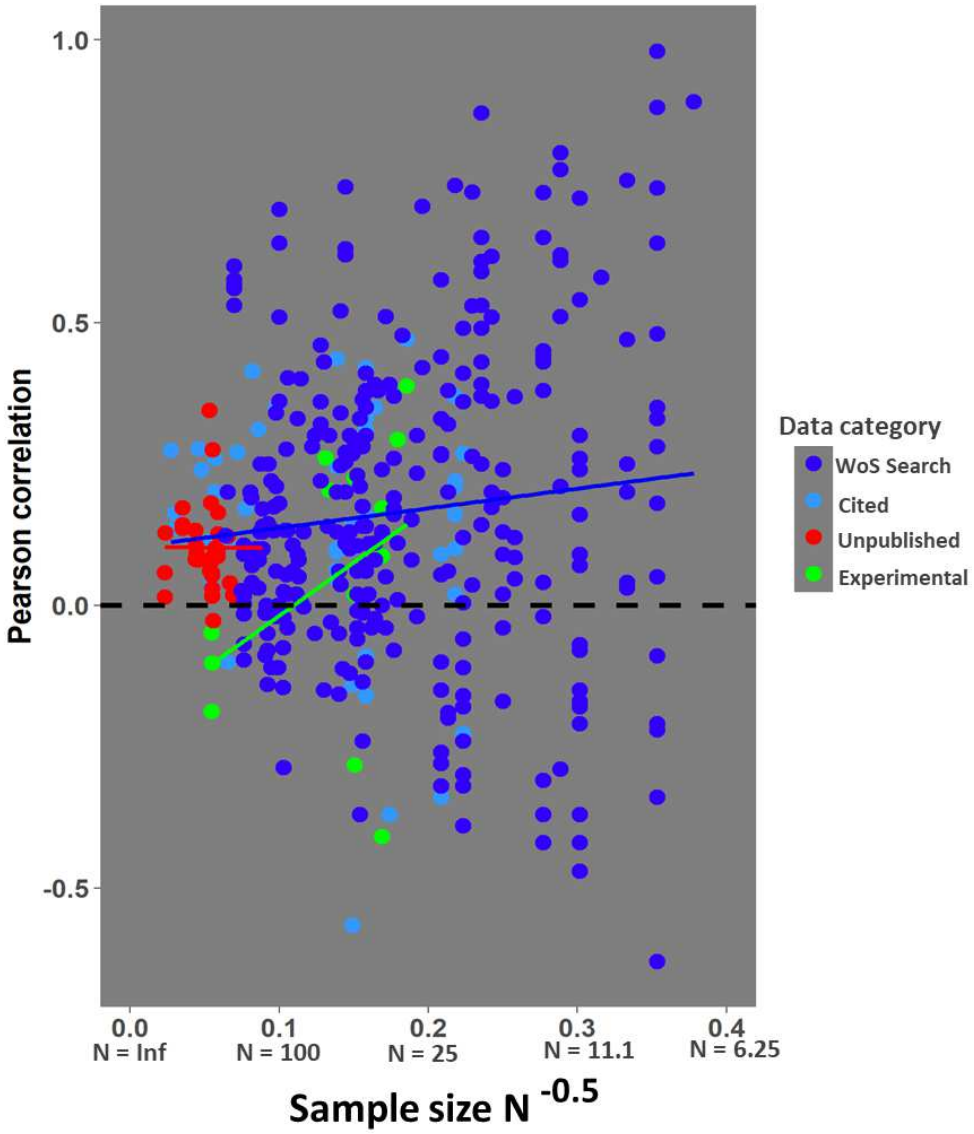

a
Age
n = 38 (5907)
Behaviour
n = 32 (6521)
Physiology
n = 9 (560)
Plumage
n = 145 (33494)
Heterozygosity
n = 9 (628)
Body condition
n = 23 (1730)
Others
n = 11 (501)
Body size
n = 357 (25456)
b
Web of Science Search
n = 302 (15971)
Cited Studies
n = 55 (9485)
Unpublished data
n = 32 (16611)
Experimental data
n = 13
(1414)
c
Site
n = 22 (34672)
n = 22 (24771)
Month
n = 32 (20729)
n = 31 (51995)
Observer
n = 32 (26542)
n = 32 (44112)
same
different
-0.5
-0.25
0
0.25
0.5
Pearson's r & 95%CI

Figure 2. Funnel plot showing single estimates of assortative mating for body size (r-values) in relation to sample size and data source. Sample size is plotted as x = N-0.5, such that infinite sample size is reached when x = 0. The regression lines refer to all data from the literature ('Web of Science Search' and 'Cited studies' together; blue), 'Unpublished data' (based on the 'nearest model'; red), and 'Experimental data" from the zebra finch study (green). The dashed line indicates no assortative mating (r = 0).



Scrutinizing assortative mating in birds

Daiping Wang, Wolfgang Forstmeier, Mihai Valcu, Niels Dingemanse, Martin Bulla, Christiaan Both, Renee Duckworth, Lynna Marie Kiere, Patrik Karell, Tomáš Albrecht, Bart Kempenaers

SUPPLEMENT

Supplementary Methods

Description of long-term field studies ('Unpublished data')

(1) Barn swallows

Barn swallows were studied in four separate breeding colonies in the Trebon area, South Bohemia, Czech Republic, between 2010 and 2015 (six breeding seasons). All birds were captured during the early breeding season and wing length, tarsus length and body mass measured. Right and left tail streamer lengths were measured to the nearest mm, and we use the average of the two measures as tail length. Each individual received an aluminium ring (National Museum Prague) and a unique combination of plastic colour rings (AVINET) before release. Phenotypic (morphological) measurements were taken early in the season, while members of social pairs were identified later in the season by the colour band combination of individuals that incubated or provisioned offspring at active nests. Nests were checked daily to determine the onset of egg laying. In the analysis, we only included first breeding attempts of each social pair in each year. For further details see Petrzelkova et al. (2015) and Wilkins et al. (2016).

(2) Blue-footed boobies

Blue-footed boobies were studied at the Isla Isabel colony off the Pacific coast of Mexico. Since 1988, reproduction has been monitored each year by marking nests, recording nest contents, and banding nestlings between February and July, and >90% of the breeders in the study area were banded with a unique number(Drummond et al. 2003). Between 2010 and 2013, culmen, ulna, and body mass were measured for a total of 551 pairs (510 unique pairs). This sample comprised two subsamples: (1) 170 pairs measured between December and March before egg laying; these pairs were defined based on behaviours including mutual courting, allopreening, and joint territory defence over 4-5 days of behavioural observations prior to capture; (2) 381 pairs measured between February and April 2011 when their broods were 10-40 days old. For further details see Kiere et al. (2016).

(3) Blue tits: study site Kolbeterberg

A population of blue tits was studied in a 35 ha plot of mixed deciduous woodland in Vienna, Austria (48°139 N, 16°209 E). The forest is dominated by oak (*Quercus robur*), beech (*Fagus sylvatica*) and ash (*Fraxinus excelsior*) and contained maximally 220 nest-boxes. We captured blue tits in their nestbox, either in winter while they were roosting or in late spring during nestling feeding. Unbanded birds were marked with a unique combination of plastic colour bands and a numbered metal ring. At capture, we measured tarsus and wing length with a calliper to the nearest 0.05 mm, and body mass with an electronic balance to the nearest 0.1 g. For more details see (Delhey et al. 2003; Foerster et al. 2003).

(4) Blue tits at Westerholz

The project is part of a long-term study on the breeding biology of blue tits, conducted in a mixed deciduous/coniferous woodland ('Westerholz', 48°08´26´´, N 10°53´29´´E) near Landsberg am Lech, southern Germany. The study area is an unmanaged part of the forest ('Reiherschlag', ca. 40 ha), which is dominated by mature oak trees and contains 277 nestboxes (since 2007) with 60-100 breeding attempts of blue tits each year. All breeding pairs were captured inside the nestbox, either in the winter preceding the breeding season (roosting), or during the breeding season (when adults fed 8-10-day-old nestlings, using an automated nestbox trap). We marked them with a unique combination of colour bands, took a small blood sample from the brachial vein (approximately 50 ml) for later parentage analysis, and measured tarsus, wing length and body mass. For more details see (Schlicht et al. 2012).

(5) Great tits

The studied population of great tits breeding in nest boxes is in Southern Germany (Bavarian Landkreis Starnberg; 47°58´N, 11°14´ E). The nest boxes were located in 12 plots established in 2009 with each plot approximately 9 hectares in size and consisting of a regular grid of 50 nest boxes with 50 m between adjacent boxes. Nest boxes were checked twice per week from April onward to determine lay date (back-calculated assuming that one egg was laid per day), onset of incubation and clutch size. Nestlings were blood sampled and marked with an aluminium ring when they were 6 days old. Parents were caught with a spring trap in the nest box the next day, measured, bled, and marked with a unique combination of rings if not ringed previously. For more details see (Araya-Ajoy et al. 2016).

(6) Pied flycatchers

Since 2007, breeding pairs of pied flycatchers (ca. 300) in Drenthe (NL, 52°49'N, 6°22'E) in ca. 1100 nest boxes distributed across 12 plots, 9 with 100 and 3 with ca. 50 nest boxes each. Pairs are defined as a male and a female that were caught during nestling feeding in a nest box (for over 90% of all nests the female identity was known and male identity was known for ca 85%). Polygyny is rather rare in this population (<4% in most years). We measured tarsus length (to the nearest 0.1 mm), the length of the third primary (from outside, to the nearest 0.5 mm) and body weight (to the nearest 0.1 g) of all birds upon capture. Several observers were measuring the birds during each year, and it was mostly the same observer measuring the male and female of a pair. Females were also caught (if possible) during incubation (around day 7 after clutch completion) and at this moment the females are considerably heavier than during nestling feeding. We did not always aim catching females again during nestling feeding if we knew their identity. For more details see (Both et al. 2017).

(7) Semipalmated sandpipers

The study area of this population of Semipalmated sandpipers is located near Barrow, Alaska (71° 32´N, 156°65´W). Breeding adults were marked with an aluminium US Geological Survey band, a unique combination of 4 colour bands, and a green flag with embedded glass passive–integrated tag (Biomark: 9.0 mm × 2.1 mm, 0.087 g, 134.2 kHz, ISO FDXB, http://www.biomark.com/). We took a small (ca. 50 µl) blood sample from a brachial vein for molecular sexing, weighed each bird (to the nearest 0.1 g) using a digital balance, and measured tarsus, culmen, and total head (to the nearest 0.1 mm) with callipers and measured

wing length (to the nearest 0.5 mm) with a ruler. More details were provided in (Bulla et al. 2014).

(8) Tawny owls

Tawny owls were studied in a nest box equipped study area of ca. 250 km^2 in southern Finland (60° 15’ N, 24° 15’ E) between 1978 and 2015. Throughout the study period nearly all pairs nested in nest boxes, which were provided in high abundance. Each year starting in mid-April, all boxes and other possible breeding sites were checked. Practically all females and males were trapped when the offspring were 1–2 weeks old. Brooding females were taken from their nest boxes in the evening by netting them at the opening of the nest box. After handling, the female was put back into the nest box and a swing-door trap for the male was mounted in front of it and left over night. In the following morning, traps were checked and the males were handled. During handling the parental birds were ringed (if unbanded) and their wing length and tail length were measured with a ruler and body mass was measured with a spring scale.

In this data set the definition of a pair is when both the male and the female has been caught and identified in the same breeding occasion. Tawny owls breed only once during a breeding season and do not re-nest if the breeding fails or the brood is depredated. The frequency of extra-pair young is low in tawny owls and estimated to 2.7 % in Saladin et al. (2007). More information on the study population and morphological traits were provided in (Karell et al. 2009; Brommer et al. 2015).

(9) Western Bluebirds

Data of Western bluebirds were collected over 15 breeding seasons (2001–2015) from a nest-box population of Western bluebirds in western Montana, USA (see Duckworth, 2006 for study site details). GPS coordinates for all nest boxes were recorded each year. Each year, nest boxes were visited at least twice weekly during the breeding season (April–August) to monitor nest progress, to determine the affiliation of breeding pairs with specific boxes, and to band offspring and adults. Adults were captured at each site using traps baited with mealworms to mark them with a unique colour band combination, and take standard morphological measurements, including body mass and length of the tarsus, tail, wing, and bill (for details on morphological variation see Duckworth and Semenov 2017). Individuals were identified as a breeding pair if they were observed together defending a territory and nest box and jointly participating in breeding activities (courtship feeding of female by male, male feeding female on nest, both parents feeding nestlings).

Description of 'Experimental data'

Morphological measurements of zebra finches

All birds of the domesticated population (experiments 1-3 below) were measured by the same observer (W.F.) for body mass (to the nearest 0.1g) using electronic scales, for wing length (to the nearest 0.5mm) using a wing ruler, and for tarsus length (to the nearest 0.1mm) using a wing ruler, when they reached 100-120 days of age (prior to release into the experimental aviaries). All birds of the wild-derived population (experiments 4-5 below) were measured by Malika Ihle for body mass (to the nearest 0.1g) using electronic scales on the day of their release into the experimental aviaries (when reaching 45 days of age). Measurements of their

tarsus length (to the nearest 0.1mm) using a wing ruler were all taken by Ulrich Knief (between 25-04-2012 and 04-05-2012) after the birds had formed pair bonds (when birds were 284 ± 46 days old, range 190 – 378 days). Note that the latter tarsus measurements violate the criterion of measuring before pair formation (hence the marking by asterisks in Table S6), yet we assume that tarsi are fully grown by 45 days of age and do not change thereafter.

Observations of pair bonds in 5 experimental studies

(1) Domesticated population: inbreeding avoidance study 2007

This experiment was designed to test whether cross-fostered zebra finches avoid pairing with unfamiliar genetic full-sibs (following up on Schielzeth et al. 2008). The studied domesticated population was kept at the Max Planck Institute for Ornithology in Seewiesen, Germany since 2004 (population # 18 in Forstmeier et al. 2007). Housing conditions, diet and aviary specifications for breeding have been described in detail in the Supplementary File to Wang et al. (2017). In this study, we used 36 males and 36 females that originated from 12 families (always 3 sons and 3 daughters that were all unfamiliar from each family). We used 6 experimental aviaries, each equipped with 6 nest boxes, and in each we released the members of two families (6 males and 6 females) to observe to which extent pair bonds form within and between families. The experiment lasted for 12 weeks (11-09-2007 to 03-12-2007). All birds were colour-banded for individual recognition (like in all following experiments). Observations of pair bonding behaviours (allopreening, sitting in body contact, and visiting a nest-box together) were carried out at least once per day, but around 6-8 times a day at the beginning of the experiment. We defined the start of one pair bond as the time when the female did not

147 show any pair bonding behaviour anymore with another male. The end of a pair bond was
148 defined by either the first observation of another exclusive pair bond (if applicable) or the last
149 observation of pair bonding behaviour (if the pair bond did not seem to last until the end of the
150 experiment). Some individuals engaged in multiple pair bonds, either sequentially (considered
151 as monogamous) or simultaneously (polygamous). For this present analysis we only included
152 monogamous pairs bonds that had been observed (n = 44).

153 *(2) Domesticated population: inbreeding depression study 2009*

154 This experiment was similar to the previous one, but it comprised the inbred and outbred
155 offspring that had been produced during the previous experiment. Each of the 6 aviaries again
156 received 6 males and 6 females (half inbred (F = 0.25), and half outbred (F = 0)) that were all
157 unfamiliar. The experiment lasted 16 weeks (07-04-2009 to 28-07-2009). Following daily
158 observations, 35 monogamous pair bonds were formed.

159 *(3) Domesticated population: selection lines 2014/15*

160 The details of this experiment have been described in Wang et al. (2017). Briefly, the birds are
161 from the same captive population as described above. In 2009 we initiated the breeding of lines
162 that were selected for high versus low breeding values for male courtship rate (two high lines,
163 two unselected control lines, two low lines; see Mathot et al. (2013)). The third generation of
164 these six lines consisted of a total of 343 females and 338 males. A subset of 219 females and
165 217 males (about equally representing the six lines) were randomly divided into 4 cohorts that
166 were tested sequentially due to the limited number of aviaries (n = 9) available. Each cohort
167 went through two rounds of breeding, in each of which they encountered a different set of

potential partners over a 7 week period. During each breeding round of a cohort, we carried out daily observations as described above. Observations lasted approximately 30 min (total across the nine aviaries) and were carried out approximately 120 times per breeding round. Across the four cohorts and the two breeding rounds we identified a total of 423 pair bonds within the 72 aviaries. Of these, 342 bonds were classified as monogamous (see Wang et al. 2017) and included into this study.

(4) Wild-derived population: compatibility study 2012

This wild-derived population (described as population # 4 in Forstmeier et al. 2007) was derived from wild-caught birds from northern Victoria about 12-15 generations ago. In 1992, 12 males and 12 females had been exported to Bielefeld, Germany, and bred there. In 2009, 109 individuals were transferred from Bielefeld to Seewiesen, where the population has been maintained since. All birds of the experiment hatched in the summer of 2011 in large semi-outdoor aviaries. Shortly after independence (when birds were 45 days old), they were put into 8 mixed-sex peer-groups of 10 males and 10 females. When birds reached sexual maturity (100 days old) they were colour-banded individually, and peer-groups were joined two by two (yielding four groups, each allowing for 20 possible pairs to form). Following observations for pair bond identification (as described above), 58 pairs were identified during the winter of 2011/2012 and included into this study. For more details see Ihle et al. (2015).

(5) Wild-derived population: inbreeding depression study 2012

This experiment is identical to the previous one (experiment 4), yet it comprised a balanced mix of inbred (F = 0.25) and outbred (F = 0) offspring (like in experiment 2). When reaching 45 days

of age, offspring went into four mixed-sex peer-groups (each group including five outbred males, five outbred females, five inbred males and five inbred females). When reaching about 100 days of age, the peer-groups were joined two by two in two different aviaries for the whole winter. Following observations as described above we identified 31 monogamous pairs that were included in this study.

Figure S1. Assortative mating for eight morphological measures of size from 'Unpublished data' (nine field studies, species name abbreviations see Table 2; details in Tables S2-4). The left panel shows model 1 ('random alignment model', analysed by randomly selecting a single measure from multiple measures); the centre panel shows model 2 ('average model', analysed by taking the mean of all available measures of pair members); the right panel shows model 3 ('nearest model', using the measures of pair members that were taken closest to the presumed time of pair formation). Here, the estimates of Semipalmated sandpipers' wing and Tawny owl's mass are the highest (significant positive assortment) across these three models.

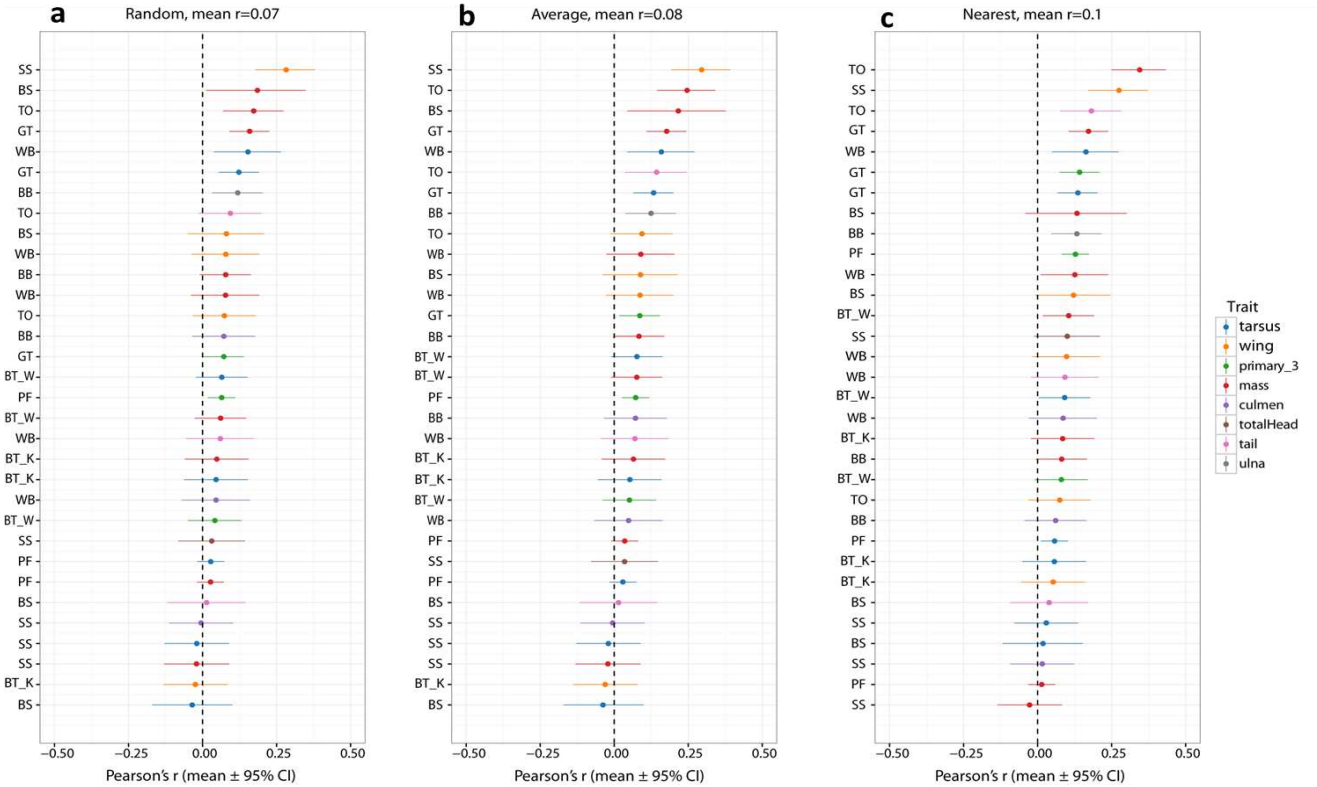


Table S1. Summary of the strength of assortative mating from literature data across eight types of traits. The mixed-effect model includes 624 estimates from 'Web of Science Search' and 'Cited studies'. Pearson correlation coefficients of assortment (weighed by sample size $(n-3)^{0.5}$, n = number of pairs) are modelled as the response variable. P-values were calculated from t-values with infinite df. The overall intercept was removed to directly show the average Pearson correlation for each trait category (fixed effect with 8 levels). The number of Pearson correlations available for each category is given as n. For the random effects, the estimates showing the proportion of variation explained (repeatability).

			95% CI			
	Sample size	Estimate	Lower	Upper	t	p
random effects:						
Study ID	158	12%				
Species ID	117	2%				
Residual		86%				
fixed effects:						
Age	38	0.409	0.328	0.490	9.98	<0.0001
Behaviour	32	0.330	0.217	0.444	5.72	<0.0001
Body condition	23	0.240	0.148	0.331	5.14	<0.0001
Body size	357	0.198	0.150	0.246	8.09	<0.0001
Heterozygosity	9	0.235	0.076	0.395	2.90	0.004
Others	11	0.223	0.077	0.368	2.99	0.003
Physiology	9	0.302	0.093	0.512	2.83	0.005
Plumage	145	0.262	0.204	0.319	8.94	<0.0001

Table S2. Assortative mating estimates from the 'random alignment model' (model 1). For each study-trait combination the average Pearson r, the average boundaries of the 95% CI, and the number of unique pairs are indicated. Asterisks mark significant ($p < 0.05$) correlations.

Study species	Trait	r	95% CI low	95% CI up	N pairs
Blue-footed booby	culmen	0.07	-0.04	0.18	339
Blue-footed booby	mass	0.08	-0.01	0.17	509
Blue-footed booby	ulna	0.12*	0.03	0.20	510
Barn bwallow	tarsus	-0.03	-0.17	0.10	209
Barn bwallow	tail	0.02	-0.12	0.15	222
Barn bwallow	wing	0.08	-0.05	0.21	233
Barn bwallow	mass	0.19*	0.01	0.35	127
Great tit	primary 3	0.07*	0.00	0.14	811
Great tit	tarsus	0.12*	0.05	0.19	809
Great tit	mass	0.16*	0.09	0.23	809
Blue tit_K	wing	-0.02	-0.13	0.08	328
Blue tit_K	tarsus	0.05	-0.06	0.15	330
Blue tit_K	mass	0.05	-0.06	0.15	331
Pied flycatcher	tarsus	0.03	-0.02	0.07	1818
Pied flycatcher	mass	0.03	-0.02	0.07	1832
Pied flycatcher	primary 3	0.06*	0.02	0.11	1789
Semipalmated sandpiper	mass	-0.02	-0.13	0.09	320
Semipalmated sandpiper	tarsus	-0.02	-0.13	0.09	325
Semipalmated sandpiper	culmen	-0.01	-0.11	0.10	325
Semipalmated sandpiper	totalHead	0.03	-0.08	0.14	302
Semipalmated sandpiper	wing	0.28*	0.18	0.38	321
Tawny owl	wing	0.07	-0.03	0.18	341
Tawny owl	tail	0.10	-0.01	0.20	335
Tawny owl	mass	0.17*	0.07	0.27	349
Western bluebirds	culmen	0.05	-0.07	0.16	288
Western bluebirds	tail	0.06	-0.05	0.18	289
Western bluebirds	wing	0.08	-0.04	0.19	290
Western bluebirds	mass	0.08	-0.04	0.19	286
Western bluebirds	tarsus	0.15*	0.04	0.26	285
Blue tit_W	primary 3	0.04	-0.05	0.13	471
Blue tit_W	mass	0.06	-0.03	0.15	509
Blue tit_W	tarsus	0.06	-0.02	0.15	503

Table S3. Assortative mating estimates from the 'average model' (model 2). For each study-trait combination the Pearson r, the boundaries of the 95% CI, and the number of unique pairs are indicated. Asterisks mark significant ($p < 0.05$) correlations. Note that 27 out of 32 correlations are higher than those from model 1 (Table S2).

Study species	Traits	r	95% CI low	95% CI up	N pairs
Blue-footed booby	culmen	0.07	-0.03	0.18	339
Blue-footed booby	mass	0.08	0.00	0.17	509
Blue-footed booby	ulna	0.12*	0.04	0.21	510
Barn swallow	tarsus	-0.04	-0.17	0.10	209
Barn swallow	tail	0.01	-0.12	0.15	222
Barn swallow	wing	0.09	-0.04	0.21	233
Barn swallow	mass	0.22*	0.04	0.38	127
Great tit	primary 3	0.09*	0.02	0.15	811
Great tit	tarsus	0.13*	0.06	0.20	809
Great tit	mass	0.18*	0.11	0.24	809
Blue tit_K	wing	-0.03	-0.14	0.08	328
Blue tit_K	tarsus	0.05	-0.06	0.16	330
Blue tit_K	mass	0.07	-0.04	0.17	331
Pied flycatcher	tarsus	0.03	-0.02	0.08	1818
Pied flycatcher	mass	0.04	-0.01	0.08	1832
Pied flycatcher	primary 3	0.07*	0.03	0.12	1789
Semipalmated sandpiper	mass	-0.02	-0.13	0.09	320
Semipalmated sandpiper	tarsus	-0.02	-0.13	0.09	325
Semipalmated sandpiper	culmen	-0.01	-0.11	0.10	325
Semipalmated sandpiper	totalHead	0.03	-0.08	0.15	302
Semipalmated sandpiper	wing	0.30*	0.19	0.39	321
Tawny owl	wing	0.09	-0.01	0.20	341
Tawny owl	tail	0.14*	0.04	0.25	335
Tawny owl	mass	0.25*	0.14	0.34	349
Western bluebirds	culmen	0.05	-0.07	0.16	288
Western bluebirds	tail	0.07	-0.05	0.18	289
Western bluebirds	wing	0.09	-0.03	0.20	290
Western bluebirds	mass	0.09	-0.03	0.20	286
Western bluebirds	tarsus	0.16*	0.04	0.27	285
Blue tit_W	primary 3	0.05	-0.04	0.14	471
Blue tit_W	mass	0.08	-0.01	0.16	509
Blue tit_W	tarsus	0.08	-0.01	0.16	503

Table S4. Assortative mating estimates from the 'nearest model' (model 3). For each study-trait combination the Pearson r, the boundaries of the 95% CI, and the number of unique pairs are indicated. Asterisks mark significant ($p < 0.05$) correlations. Note that 27 out of 32 correlations are higher than those from model 1 (Table S2).

Study species	Traits	r	95% CI low	95% CI up	N pairs
Blue-footed booby	culmen	0.06	-0.04	0.17	346
Blue-footed booby	mass	0.08	-0.01	0.17	515
Blue-footed booby	ulna	0.13*	0.05	0.22	517
Barn swallow	tarsus	0.02	-0.12	0.15	209
Barn swallow	tail	0.04	-0.09	0.17	222
Barn swallow	wing	0.12	-0.01	0.25	233
Barn swallow	mass	0.13	-0.04	0.30	127
Great tit	primary 3	0.14*	0.07	0.21	811
Great tit	tarsus	0.14*	0.07	0.20	809
Great tit	mass	0.17*	0.10	0.24	809
Blue tit_K	wing	0.05	-0.06	0.16	328
Blue tit_K	tarsus	0.06	-0.05	0.16	330
Blue tit_K	mass	0.08	-0.02	0.19	331
Pied flycatcher	tarsus	0.06*	0.01	0.10	1824
Pied flycatcher	mass	0.01	-0.03	0.06	1838
Pied flycatcher	primary 3	0.13*	0.08	0.17	1795
Semipalmated sandpiper	mass	-0.03	-0.14	0.08	322
Semipalmated sandpiper	tarsus	0.03	-0.08	0.14	329
Semipalmated sandpiper	culmen	0.02	-0.09	0.12	329
Semipalmated sandpiper	totalHead	0.10	-0.01	0.21	306
Semipalmated sandpiper	wing	0.28*	0.17	0.37	324
Tawny owl	wing	0.07	-0.03	0.18	343
Tawny owl	tail	0.18*	0.08	0.28	337
Tawny owl	mass	0.34*	0.25	0.43	351
Western bluebirds	culmen	0.09	-0.03	0.20	289
Western bluebirds	tail	0.09	-0.02	0.21	290
Western bluebirds	wing	0.10	-0.02	0.21	291
Western bluebirds	mass	0.13*	0.01	0.24	287
Western bluebirds	tarsus	0.16*	0.05	0.27	286
Blue tit_W	primary 3	0.08	-0.01	0.17	471
Blue tit_W	tarsus	0.09*	0.00	0.18	503
Blue tit_W	mass	0.10*	0.02	0.19	509

Table S5. Assortative mating estimates under different contexts (models 4-6). Here, model 4 reveals observer effect, model 5 reveals temporal autocorrelation, and model 6 reveals spatial autocorrelation (see detailed description of each model in methods section). For the three random effects we show the proportion of variance explained (repeatability). The overall intercept was removed to directly show the average degree of assortative mating and 95% CI for each of the two levels of the fixed effect and its significance in terms of t-values and p-values (calculated with infinite df). The fixed-effect level 'Same' refers to measurements from the same observer (in model 4), from the same month (in model 5), and from the same site (in model 6), while 'Different' refers to measurements from different observers (model 4), measurements taken more than 30 days apart (model 5), or measurements taken more than 10m apart (model 6).

				95% CI			
Model	Effect type	Effect	Estimate	Lower	Upper	t	p
Observers	Random	Trait (n =8)	0%				
(model 4)	(variance)	Study (n = 7)	1.3%				
		Trait × Study (n = 22)	0%				
		Residual	98.7%				
	Fixed	Same	0.075	0.036	0.114	3.45	< 0.0001
		Different	0.023	-0.016	0.062	1.01	0.31
Month	Random	Trait (n =8)	0%				
(model 5)	(variance)	Study (n = 9)	0.5%				
		Trait × Study (n =32)	0				
		Residual	99.5%				
	Fixed	Same	0.110	0.071	0.149	6.92	< 0.0001
		Different (> 30 days)	0.014	-0.025	0.053	0.97	0.33
Site	Random	Trait (n = 8)	0%				
(model 6)	(variance)	Study (n = 9)	0.6%				
		Trait × Study (n =32)	0.6%				
		Residual	98.8%				
	Fixed	Same	0.073	0.053	0.093	5.17	< 0.0001
		Different (>10m)	0.017	-0.003	0.037	1.28	0.2

Table S6. Data summary for experimental studies on captive zebra finches. Here, each correlation estimate r is the weighted (by $(n-3)^{0.5}$, with n = number of pairs) average of correlation coefficients calculated within experimental aviaries. Experiments are numbered as in the Supplementary Methods section. Tarsus length from experiments 4 and 5 (marked with asterisks) were measured after releasing the birds into the aviaries.

Experiment	Population	Trait	n pairs	n aviaries	r
1	domesticated	mass	44	6	0.01
1	domesticated	tarsus	44	6	0.25
1	domesticated	wing	44	6	-0.26
2	domesticated	mass	35	6	-0.45
2	domesticated	tarsus	35	6	0.20
2	domesticated	wing	35	6	0.07
3	domesticated	tarsus	331	67	-0.25
3	domesticated	mass	336	68	-0.10
3	domesticated	wing	336	68	-0.12
4	wild-derived	mass	31	2	0.30
4	wild-derived	tarsus*	29	2	0.38
5	wild-derived	mass	58	4	0.27
5	wild-derived	tarsus*	56	4	0.21

Table S7. Linear mixed model explaining the degree of assortative mating (402 Pearson r estimates) as a function of data source. For the three random effects we show the proportion of variance explained (repeatability). The overall intercept was removed to directly show the average degree of assortative mating and 95%CI for each of the four levels of the fixed effect and its significance in terms of t-values and p-values (calculated with infinite df). Pearson r estimates for 'Unpublished data' field studies were taken from the 'nearest model' (see model 3 in methods section).

			95% CI			
	Sample size (n)	Estimates	Lower	Upper	t	p
Random effects:						
Study	85	7%				
Species	73	0%				
Trait-type	7	0%				
Residual		93%				
Fixed effects:						
Web of Science Search	302	0.201	0.158	0.244	9.03	<0.0001
Cited studies	55	0.135	0.051	0.219	3.11	0.002
Unpublished data	32	0.106	0.010	0.202	2.17	0.030
Experimental data	13	-0.003	-0.279	0.273	-0.02	0.983

Reference of Supplement

1 Petrzelkova, A. *et al.* Brood parasitism and quasi-parasitism in the European barn swallow Hirundo rustica rustica. *Behav Ecol Sociobiol* **69**, 1405-1414, doi:10.1007/s00265-015-1953-6 (2015).

2 Wilkins, M. R. *et al.* Phenotypic differentiation is associated with divergent sexual selection among closely related barn swallow populations. *J Evolution Biol* **29**, 2410-2421, doi:10.1111/jeb.12965 (2016).

3 Drummond, H., Torres, R. & Krishnan, V. V. Buffered development: Resilience after aggressive subordination in infancy. *Am Nat* **161**, 794-807, doi:Doi 10.1086/375170 (2003).

4 Kiere, L. M., Ramos, A. G. & Drummond, H. No evidence that genetic compatibility drives extra-pair behavior in female blue-footed boobies. *J Avian Biol* **47**, 871-879, doi:10.1111/jav.01061 (2016).

5 Delhey, K., Johnsen, A., Peters, A., Andersson, S. & Kempenaers, B. Paternity analysis reveals opposing selection pressures on crown coloration in the blue tit (Parus caeruleus). *P Roy Soc B-Biol Sci* **270**, 2057-2063, doi:10.1098/rspb.2003.2460 (2003).

6 Foerster, K., Delhey, K., Johnsen, A., Lifjeld, J. T. & Kempenaers, B. Females increase offspring heterozygosity and fitness through extra-pair matings. *Nature* **425**, 714-717, doi:10.1038/nature01969 (2003).

7 Schlicht, L., Girg, A., Loes, P., Valcu, M. & Kempenaers, B. Male extrapair nestlings fledge first. *Anim Behav* **83**, 1335-1343, doi:10.1016/j.anbehav.2012.02.021 (2012).

8 Araya-Ajoy, Y. G. *et al.* Sources of (co)variation in alternative siring routes available to male great tits (Parus major). *Evolution* **70**, 2308-2321, doi:10.1111/evo.13024 (2016).

9 Both, C. *et al.* Delayed age at first breeding and experimental removals show large non-breeding surplus in Pied Flycatchers. *Ardea* **105**, 43-60, doi:10.5253/arde.v105i1.a2 (2017).

10 Bulla, M., Valcu, M., Rutten, A. L. & Kempenaers, B. Biparental incubation patterns in a high-Arctic breeding shorebird: how do pairs divide their duties? *Behav Ecol* **25**, 152-164, doi:10.1093/beheco/art098 (2014).

11 Saladin, V., Ritschard, M., Roulin, A., Bize, P. & Richner, H. Analysis of genetic parentage in the tawny owl (Strix aluco) reveals extra-pair paternity is low. *J Ornithol* **148**, 113-116, doi:10.1007/s10336-006-0109-x (2007).

12 Karell, P., Ahola, K., Karstinen, T., Zolei, A. & Brommer, J. E. Population dynamics in a cyclic environment: consequences of cyclic food abundance on tawny owl reproduction and survival. *J Anim Ecol* **78**, 1050-1062, doi:10.1111/j.1365-2656.2009.01563.x (2009).

13 Brommer, J. E., Karell, P., Aaltonen, E., Ahola, K. & Karstinen, T. Dissecting direct and indirect parental effects on reproduction in a wild bird of prey: dad affects when but not how much. *Behav Ecol Sociobiol* **69**, 293-302, doi:10.1007/s00265-014-1842-4 (2015).

14 Duckworth, R. A. Aggressive behaviour affects selection on morphology by influencing settlement patterns in a passerine bird. *P Roy Soc B-Biol Sci* **273**, 1789-1795, doi:10.1098/rspb.2006.3517 (2006).

15 Duckworth, R. A. & Semenov, G. A. Hybridization Associated with Cycles of Ecological Succession in a Passerine Bird. *Am Nat* **190**, E94-E105, doi:10.1086/693160 (2017).

16 Forstmeier, W., Segelbacher, G., Mueller, J. C. & Kempenaers, B. Genetic variation and differentiation in captive and wild zebra finches (Taeniopygia guttata). *Mol Ecol* **16**, 4039-4050, doi:10.1111/j.1365-294X.2007.03444.x (2007).

17 Schielzeth, H., Burger, C., Bolund, E. & Forstmeier, W. Sexual imprinting on continuous variation: do female zebra finches prefer or avoid unfamiliar sons of their foster parents? J Evolution Biol 21, 1274-1280, doi:10.1111/j.1420-9101.2008.01568.x (2008).
18 Wang, D., Kempenaers, N., Kempenaers, B. & Forstmeier, W. Male zebra finches have limited ability to identify high-fecundity females. *Behav Ecol* **28**, 784-792 (2017).
19 Mathot, K. J., Martin, K., Kempenaers, B. & Forstmeier, W. Basal metabolic rate can evolve independently of morphological and behavioural traits. *Heredity* **111**, 175-181, doi:10.1038/hdy.2013.35 (2013).
20 Wang, D., Forstmeier, W. & Kempenaers, B. No mutual mate choice for quality in zebra finches: Time to question a widely held assumption. *Evolution* **71**, 2661-2676 (2017).
21 Ihle, M., Kempenaers, B. & Forstmeier, W. Fitness Benefits of Mate Choice for Compatibility in a Socially Monogamous Species. *Plos Biol* **13**, e1002248, doi:10.1371/journal.pbio.1002248 (2015).

Chapter 5: Genetic constraints of female promiscuity: male corollary or independent trajectory?

Short title: Genetic constraints of female promiscuity

Abstract: The question of why females of many socially monogamous species engage in copulations outside the social pair bond has intrigued behavioral ecologists for many decades, especially because the benefits of such promiscuous behavior often do not seem to outweigh the costs. Hence, models of genetic constraint have been proposed, where female promiscuity emerges as a genetic corollary of alleles that are either beneficial for male extra-pair mating success (intersexual pleiotropy hypothesis) or beneficial for female fecundity (intrasexual pleiotropy hypothesis). In a first empirical test using captive zebra finches we had found support for the former hypothesis, suggesting that artificial selection on male sex drive could alter female extra-pair mating behavior as a genetic corollary. Here, we directly follow up on this suggestion and re-examine both hypotheses after establishing selection lines for male sex drive. After testing for intersexual pleiotropy with much increased statistical power, we now have to revise our previous conclusions, because the new data does not confirm the idea that male and female promiscuity are genetically homologous traits. However, we find some support for the idea that female promiscuity is genetically correlated with female fecundity, calling for more empirical tests of the intrasexual pleiotropy hypothesis. We also find that female extra-pair mating behavior is strongly context dependent, rendering genetic studies difficult and suggesting that social network analyses might shed more light on when and why females mate outside the pair bond.

Prepared as: Daiping Wang, Wolfgang Forstmeier, Katrin Martin, Alastair Wilson and Bart Kempenaers: Genetic constraints of female promiscuity: male corollary or independent trajectory?

Genetic constraints of female promiscuity: male corollary or independent trajectory?

Daiping Wang[1], Wolfgang Forstmeier[1], Katrin Martin[1], Alastair Wilson[2] and Bart Kempenaers[1]

[1] Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany

[2] Centre of Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Penryn Campus, Penryn, Conrwall, TR10 9FE, UK

Short title: Genetic constraints of female promiscuity

Keywords: quantitative genetics, promiscuity, female EPP, selection lines, fecundity

Figures & Tables: 5 figures, 1 table

Supplementary materials: 18 tables

Address for correspondence:

Wolfgang Forstmeier, Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Str. 7, 82319 Seewiesen, Germany, Phone: 0049-8157-932346

Email: forstmeier@orn.mpg.de

Abstract

The question of why females of many socially monogamous species engage in copulations outside the social pair bond has intrigued behavioral ecologists for many decades, especially because the benefits of such promiscuous behavior often do not seem to outweigh the costs. Hence, models of genetic constraint have been proposed, where female promiscuity emerges as a genetic corollary of alleles that are either beneficial for male extra-pair mating success (intersexual pleiotropy hypothesis) or beneficial for female fecundity (intrasexual pleiotropy hypothesis). In a first empirical test using captive zebra finches we had found support for the former hypothesis, suggesting that artificial selection on male sex drive could alter female extra-pair mating behavior as a genetic corollary. Here, we directly follow up on this suggestion and re-examine both hypotheses after establishing selection lines for male sex drive. After testing for intersexual pleiotropy with much increased statistical power, we now have to revise our previous conclusions, because the new data does not confirm the idea that male and female promiscuity are genetically homologous traits. However, we find some support for the idea that female promiscuity is genetically correlated with female fecundity, calling for more empirical tests of the intrasexual pleiotropy hypothesis. We also find that female extra-pair mating behavior is strongly context dependent, rendering genetic studies difficult and suggesting that social network analyses might shed more light on when and why females mate outside the pair bond.

Key words: quantitative genetics, promiscuity, female EPP, selection lines, fecundity

Introduction

Explaining why females in socially monogamous species actively engage in mating outside the pair bond has intrigued behavioural ecologists for many decades (Petrie and Kempenaers 1998; Griffith et al. 2002; Westneat and Stewart 2003; Forstmeier et al. 2014; Maldonado-Chaparro et al. 2018) (1-5). Mating outside the pair bond seems obviously adaptive for males because additional offspring mean higher fitness (Albrecht et al. 2007; Webster et al. 2007) (6, 7). However, the frequently observed female promiscuity in monogamous species is puzzling: it does not increase the number of offspring that females can produce and even may bring about additional costs such as predation risk, sexually transmitted diseases, withdrawal of paternal care and punishment by social mate (Forstmeier et al. 2014) (4). In birds, more than 90% of species breed in socially monogamous pairs and female extra-pair mating behavior is often found in these species (Griffith et al. 2002; Sheldon and Mangel 2014) (2, 8). Hence, birds have served as paragons of studying the evolution of female promiscuity. For more than two decades, the majority of research explaining the occurrence female extra-pair mating behavior has been conducted under the framework of adaptation highlighting the potential benefits (Petrie and Kempenaers 1998; Griffith et al. 2002; Hsu et al. 2015) (1, 2, 9). The proposed benefits could be either indirect genetic (Fox and Rauter 2003; Kempenaers 2007; Szulkin et al. 2013) (10-12) or direct ecological (Heg et al. 1993; Lombardo and Thorpe 2000; Sheldon & Mangel 2014) (8, 13, 14). Yet, despite much empirical work, the general support for adaptive scenarios is rather limited (Schmoll et al. 2009; Sardell et al. 2012; Hsu et al. 2014; Forstmeier et al. 2014) (4, 15-17). Therefore, alternative non-adaptive explanations might deserve special attention (Hsu et al. 2015) (9).

Taking the perspective of quantitative genetics, several hypotheses of 'genetic constraint' have been proposed to solve this evolutionary puzzle of apparent non-adaptation (Halliday and Arnold 1987; Kirkpatrick and Barton 1997; Arnqvist and Kirkpatrick 2005) (18-20). These hypotheses state that the alleles, causing female promiscuity, have additional pleiotropic effects that are beneficial and, hence, maintain the alleles in the population. Depending on whether the pleiotropic effect is expressed in males or females, we distinguish two types of hypotheses.

(1) The hypothesis of 'intersexual pleiotropy' proposes that female and male promiscuity might be homologous traits that are affected by the same sets of genes (Halliday and Arnold 1987) (18). Alleles that increase male promiscuity can be positively selected and maintained in the population, and these alleles, when inherited to a daughter, might cause female promiscuity as a by-product (i.e. 'male corollary') even if promiscuity is not adaptive for females. To test this

hypothesis, one needs to examine whether female promiscuity is genetically correlated with measures of male promiscuity (cross-sex genetic covariance).

(2) The hypothesis of 'intrasexual pleiotropy' argues that the maintenance of female promiscuity is because its causal alleles have pleiotropic effects on other female traits ('female independent trajectory') that are positively selected (Arnqvist and Kirkpatrick 2005; Forstmeier 2007) (20, 21). Female responsiveness to male courtship might be genetically linked to female fecundity, because courtship may proximately stimulate egg production (Bolund et al. 2012) (22). Alternatively, genetic variants underlying female sexual responsiveness towards her social mate may be favored by selection because frigidity can lead to infertility and reduced fitness (Arnqvist and Kirkpatrick 2005) (20). Such positively selected alleles for responsiveness towards the social mate could increase female responsiveness towards extra-pair males as well. To test this hypothesis, one needs to examine whether female promiscuity is genetically correlated to either female fecundity or to female responsiveness towards her social mate.

There has been little empirical work on non-adaptive hypotheses, partly because of the dominance of adaptive explanations and partly because of extensive data requirement for quantitative genetic models (Forstmeier et al. 2011; Forstmeier et al. 2014) (4, 23). Empirical testing of the hypotheses using field data on extra-pair paternity has been hindered by the low levels of heritability of male and female promiscuity (Reid et al. 2011; Reid 2012; Reid et al. 2014; Wilson and Poissant 2016) (24-27). The main problem is that realized patterns of paternity depend on many factors other than intrinsic inclination to seek extra-pair copulation (e.g. mating preferences, sperm competition, mate guarding).

In an earlier study (Forstmeier et al. 2011) (23) we tried to overcome these difficulties by using captive zebra finches which allowed us to supplement the data on realized levels of extra-pair paternity with detailed observations on behaviors that reflect an individual's propensity of engaging in extra-pair mating. In that study we found clear support for the hypothesis of 'intersexual pleiotropy' (male and female promiscuity being homologous traits) and we rejected the idea of 'intrasexual pleiotropy' (responsiveness to the partner and responsiveness to extra-pair males being independent traits). Hence, this first empirical assessment of the two hypotheses suggested that female promiscuity could be changed indirectly by artificially selecting males for increased or reduced sex drive (measured as courtship rate, a genetic correlate of male extra-pair siring success).

In the present study, we directly follow up on that result. Using the birds of the initial study we set up artificial selection lines that were bred to either increase (two replicate high lines) or decrease (two low lines) male courtship rate, or to serve as controls (two unselected control

lines). By increasing the genetic variance in male courtship rate, we are now able to test with much increased statistical power whether female extra-pair mating behavior is genetically linked to male courtship rate and hence whether female promiscuity was changed indirectly by selection imposed on male behavior only.

Moreover, we also amend a major weakness of the initial study: zebra finches form monogamous pair bonds that usually last until one of the pair members dies. Hence, in the initial study, the behavior of a female had been assessed usually only once, i.e. in the context of being paired to the partner that she chose in one experiment. The observed behavior of a female was then assumed to be representative for that female, but alternatively it might have been more a property of the female's social environment (social pair bond, available extra-pair males) than a property of the female. To resolve this uncertainty, we here study every female with two successive partners. This allows us to better tease apart the component that is intrinsic to the female from other components. In other words, we here first examine the repeatability of female promiscuity across two social partners before quantifying its heritability and genetic covariance with other traits.

To examine the hypothesis of 'intersexual pleiotropy, we quantify whether female promiscuity is positively genetically correlated with two measures of male sexual behavior, namely (1) male courtship rate which had been under artificial selection by us, and (2) male success in siring extra-pair eggs. To examine the alternative hypothesis of 'intrasexual pleiotropy', we test whether female promiscuity is positively correlated with (3) responsiveness towards the social mate, and (4) measures of total female fecundity.

Results

Selection lines for male courtship rate

A total of six selection lines were established and bred over three consecutive generations: two lines selected for high sex drive, two for low sex drive, and two unselected control lines. Figure 1 shows, for each generation, the phenotypes (courtship rate) of all male offspring that were bred, as function of the mean breeding value of their parents (breeding values are predictions of offspring phenotypes made by a genetic model that is based on observed phenotypes of parents and their relatives, here still excluding the offspring). Reassuringly, the slope of the regression lines is close to unity, indicating that the offspring generations behaved as predicted by the genetic model. With each generation we were able to choose parents with even more extreme breeding values, which is reflected by the outward movement of high and low lines along the x-axis over progressive generations. In consequence, the offspring phenotypes became

progressively differentiated along the y-axis between the selection lines (i.e. the data points move outwards approximately following the line with a slope of unity). After three generations of selection, the average difference between the high and the low lines (in generation 'S3') reached 2.4 phenotypic standard deviations (Cohen's d (28) (Cohen 1988)). The two replicates of each type of line behaved almost identically (see Table S3) so they are not distinguished visually in Figure 1.

Apparent indirect response to selection

In order to assess whether the successful selection on male courtship rate had resulted in correlated changes in levels of extra-pair paternity in both sexes, we quantified for each individual the proportion of paternity that was outside the pair bond, when mixed flocks (containing all types of selection lines) were breeding in communal aviaries.

Altogether 190 females produced 2,951 fertile eggs during the time they were monogamously paired, 726 of which (24.6%) were sired by extra-pair males. Levels extra-pair paternity (seen from the female perspective) ranged from 37.4% in line 'high 1' to 15.8% in line 'low 2', with the other four lines showing intermediate levels (Figure 2). A statistical analysis of individual levels of extra-pair paternity where the predictor of interest, the selection regime, was coded as a continuous variable (1df; low = -1, control = 0, high = 1) suggested a significant effect ($\beta = 0.698$, $z = 3.1$, $p = 0.002$, $n = 190$, Table S4), yet note that random effect of line (6 levels) explained none of the remaining variance (Table S4), thereby failing to effectively control for pseudoreplication (females within a line are genetically related and hence non-independent).

Analyzing the paternity data from the male perspective, 188 males sired 3,067 eggs during the time that they were socially paired, 851 of which (27.7%) had been laid by females other than their social mate. The corresponding levels of extra-pair paternity ranged from 32.2% in line 'control 2' to 16.7% in line 'low 2' (Figure 2). Here, the continuous predictor of selection regime showed a non-significant trend in the expected direction ($\beta = 0.278$, $z = 1.7$, $p = 0.09$, $n = 188$, Table S5), yet again the random effect of line failed to control for non-independence (Table S5).

Repeatability of female promiscuity across two social bonds

Figure 3 illustrates the extent to which measurements of individual female promiscuity (average responsiveness towards extra-pair males and levels of extra-pair paternity) are repeatable between two breeding rounds with different partners and different sets of extra-pair males. Specifically, weighted ordinary least square regression lines indicate correlations of 0.37 (n = 151 females) and 0.24 (n = 135 females), respectively (Figure 3). Accordingly, it can also be seen from the respective permanent environment animal models (Tables S6, S7, S10 to S15), that the random effect of social pair ('Pair ID') explained considerably more variance in measures of female promiscuity than the random effects that represent female identity ('Genetic' + 'Permanent environment'). In other words, a female's level of promiscuity is a lot more consistent within a given context (social pair bond, set of extra-pair males) than between different contexts, thereby impeding the estimation of an individual female's intrinsic phenotype.

Testing the 'intersexual pleiotropy' hypothesis

Figure 4a,b illustrates the initial raw data from Forstmeier et al. (2011) that led to the suggestion that females that carry alleles for high male courtship rate (female breeding values) show an increased responsiveness to extra-pair males courting them (Fig. 4a) and higher levels of extra-pair paternity (Fig. 4b). The new data from the three types of selection lines is shown for comparison in the panels underneath (Fig. 4c and 4d). The artificially increased range in breeding values (thanks to selection lines) allows for more powerful tests, yet the indicated regression slopes turn out much shallower than suggested by the initial data. Note that these regression lines are merely for illustration, since they do not account for other influential fixed effects. The decisive tests for whether measures of male and female promiscuity are genetically correlated are presented in Table 1, where we contrast estimates of between-sex genetic correlations from 5-trait animal models based on the initial data (Table S10, S11) to those from models on the new data from selection lines (Tables S12, S13). According to the new data, between-sex genetic correlations were very close to zero when regarding the male trait for which we had artificially increased the genetic variance (courtship rate, mean of four estimates r_A = 0.04), and the trend was even opposite to expectations when regarding male extra-pair siring success (mean r_A = -0.34). These estimates stand in strong contrast to the generally positive estimates derived from the initial data (Table 1). An updated matrix of genetic correlations estimated from the joint data (initial plus selection lines) is presented in Figure 5a (summary of Tables S6 to S9 showing medians of estimates from four types of models). In this

summary, between-sex genetic correlations are weakly positive but not significantly larger than zero.

Testing the 'intrasexual pleiotropy' hypothesis

Estimates of genetic correlations between traits within the female sex are shown in Figure 5b (medians across four animal models, Table S14 to S17, based on all available data). We found that female responsiveness to extra-pair males is only weakly positively correlated (r_A = 0.26 ± 0.19) to responsiveness to the own partner, yet a bit more strongly to our measure of female fecundity (r_A = 0.41 ± 0.17), thereby providing some tentative support for the hypothesis of intrasexual pleiotropy.

Discussion

Overall, our data were more supportive of the 'intrasexual pleiotropy' hypothesis than the hypothesis of 'intersexual pleiotropy', suggesting female promiscuity is a 'female independent trait' rather than a 'male corollary' (29, 30) (Sgro et al. 1998; Eady et al. 2000).

The breeding of selection lines for male sex drive was very effective in maximizing the statistical power for testing whether measures of female promiscuity are indeed genetically correlated with male sex drive (see the increased data range in Fig. 4). The most decisive test for such a genetic correlation yielded a clear answer leading to rejection of the 'intersexual pleiotropy' hypothesis (see 'new data' in Table 1). This conclusion is not much affected by weak trends in the phenotypic data (Figs. 2 and 4d) that appear to be in line with the 'intersexual pleiotropy' hypothesis. In the case of Figure 2, statistical testing even suggested a significant effect of selection regime on female levels of extra-pair paternity, mostly stemming from reduced levels of extra-pair paternity in females from the two 'low lines'. However, the corresponding model (Table S4) failed to effectively control for pseudoreplication by specifying 'line ID' as a random effect (line ID happened to explain zero variance), which can easily happen when the total number of lines is low (here 6 lines in total). In this case, animal models that control for all non-independence of individuals via genetic relatedness (Table 1) should produce a more trustworthy answer, which we base our conclusion on. Taken together, we think that the present study effectively rejects the 'intersexual pleiotropy' hypothesis as a main explanation, despite some weak remaining correlations in the joint analysis (Figure 5a) that incorporates both the initial and the follow-up data.

More promising, in terms of explaining the maintenance of female promiscuity, is our finding of positive genetic covariance between female extra-pair responsiveness and female fecundity

(Figure 5b). We think that this finding deserves more study and should be readily addressable also in populations breeding in the wild (see below). Finally, our study reveals a great deal of context dependence of female extra-pair mating, which depended most strongly on 'Pair ID' in our quantitative genetic analyses (Figure 3, Tables S6, S7). This could be either a matter of the quality of the social pair bond, or a matter of the set of available extra-pair males, a question that calls for more detailed analyses of extra-pair mating in relation to social network characteristics (5)(Maldonado-Chaparro et al. 2018).

Explaining the discrepancy in findings with the initial study

Quantifying, for the first time, individual female extra-pair behavior across two social pair bonds (Figure 3), we have learnt that it may be dangerous to equate a female's phenotype in a single context with her overall intrinsic phenotype. This means that phenotypes used in the initial study (y-axes of Fig. 4a and 4b) contain a greater amount of noise than the phenotypes in the follow-up study (y-axes of Fig. 4c and 4d; noise should be reduced to about half by averaging among two contexts), thereby increasing the risk of obtaining a spurious positive as opposed to true positive result in the initial study. Moreover, estimates of breeding values in the initial study (x-axes of Fig. 4a and 4b) were based on fewer male relatives (N = 800) than we had after the breeding of selection lines (N = 1,651), meaning higher error along the x-axis as well. Interestingly, when updating Figures 4a and 4b with additional information on courtship rate (reducing the error in the x-axis while leaving the values on the y-axes unchanged), we obtain shallower slopes of regression lines (Fig. 4a: $\beta = 0.14$, Fig. 4b: $\beta = 0.07$), which also hints towards measurement error being responsible for a false-positive result. Finally, the initial study was based on a very limited sample of individuals (about 150 females) and it is possible that founder effects (31)(Swallow et al. 1998) resulted in some linkage disequilibrium between alleles for male and female promiscuity by chance alone. Such non-physical linkage may then have gotten broken up during the subsequent breeding of selection lines.

In conclusion, we currently consider the positive findings in the initial study (Forstmeier et al. 2011) as a classical false-positive finding that resulted from limited and relatively noisy data, but not from inadequate modelling. We have updated the calculations of the earlier models by also including clutch identity and pair identity as additional random effects, but this did not alter the conclusions that emerge from the initial data (see Table 1 and Tables S10 and S11). What seems noteworthy is that Bayesian models in MCMCglmm often gave more conservative estimates with larger standard errors than REML models in VCE, and the former estimates proved to be closer to reality in our follow-up study. This experience confirms the general notion that the

estimation of genetic correlations is fraught with difficulty when heritabilities are not as high (32) and when sample sizes are limited because phenotyping is very labor intensive.

Future directions

We found that female fecundity showed positive genetic covariance with measures of female promiscuity (Fig. 5a). This result might explain the persistence of extra-pair mating and hence may be worth following up in studies in the wild. Reid et al. (2012) (26) found positive genetic covariance between female levels of extra-pair paternity and female annual reproductive success, so it would be interesting to know whether this was due to variation in fecundity or variation in rearing success. Where quantitative genetic analyses are not feasible because detailed pedigree information is not available, one could still examine whether there is a positive phenotypic correlation between clutch size and levels of extra-pair paternity. Such analyses should focus on the proportion of eggs in a clutch that are extra-pair (rather than on the presence vs. absence of extra-pair paternity in a clutch, because the probability of detecting extra-pair mating naturally increases with the number of eggs examined). Also, such field studies may want to control for breeding density (availability of extra-pair males) as a possible confound, because both clutch size and breeding density may vary with habitat quality.

Our new analyses of female extra-pair behavior across two social environments (Figure 3) revealed a substantial amount of context-dependence of this behavior. When considering levels of extra-pair paternity, the most influential factor was the identity of the social pair ('Pair ID' in Tables S6, S7), indicating substantial consistency across multiple clutches with the same partner and much flexibility between the two social partners (Figure 3b). Such consistency at the level of the social pair rather than at the level of the female is consistent with similar findings on coal tits by (33) (Dietrich et al. 2004) and might suggest that there is variation in the strength of the social pair bond affecting paternity levels (behavioral compatibility of mates as suggested by Ihle et al. 2015) (34). When considering a female's responsiveness towards courting extra-pair males, the most influential factor was again the combination of male and female identities, i.e. who encountered whom (coded as 'Pair ID' in Tables S6, S7). Hence the occurrence of promiscuous behavior appears to depend most strongly on aspects of compatibility between individuals. The dependence on the social context might either be mostly a matter of the quality of the social pair bond or mostly a matter of the availability of specific extra-pair males. Which of these two aspects of the social environment is more important for determining extra-pair paternity levels, could be either addressed in specifically designed experiments or by targeted social network analyses as suggested by (5) (Maldonado-Chaparro et al. 2018).

Conclusions

Even though our selection experiment did not show that levels of female promiscuity can be altered by artificially selecting on male sexual behavior alone, models of genetic constraint in general remain a viable explanation for the persistence of female extra-pair mating. All examined genetic correlations in Figure 5 are positive (rather than half positive, half negative as expected from randomness) after incorporating all available data from our study population (initial study plus verification study), and following up on some of these constraints appears both promising and feasible.

Methods

Subjects

Study subjects are from the same population as described in previous studies (Forstmeier et al. 2011; Wang et al. 2017a; Wang et al. 2017b) (23, 35, 36). This population has been maintained at the Max Planck Institute for Ornithology in Seewiesen, Germany since 2004, (population # 18 in Forstmeier et al. 2007) (37). Housing conditions, diet and aviary specifications for breeding have been described in detail in the Supplementary File to (Wang et al. 2017a) (35). For the present study, the pedigree of this population comprises eight generations: Parental, F1 to F4, and four generations of selection lines (S1 to S3, see below).

Behavioural Observations

We measured behavioral traits related to the extra-pair mating under two experimental set-ups: cage experiments and aviary breeding experiments. In the cage experiments, where measurements are more standardized leading to high individual repeatability, we measured for all males in our population 'male courtship rate' towards bachelor females (the trait subjected to artificial selection) and for all females in our population 'female unpaired response' to the courtship by bachelor males (details see below). In the aviary breeding experiments, we measured, for a subset of individuals, female responsiveness to the courtship either by her social partner ('within-pair response') or by other males ('extra-pair response'). The set-up of aviary breeding experiments is more natural and more complex, so we aimed for a high number of observations per individual to make up for lower repeatabilities of behaviors.

a) Cage Experiments on Bachelor Birds

Details of arranged male-female encounters in a cage were described in (Forstmeier et al. 2011) (23). In short, each encounter is a five-minute trial including a bachelor male and a bachelor female unfamiliar to each other. For each trial we recorded the total duration (in seconds) of

male courtship: that is, song directed toward the female (referred to as 'male courtship rate'). The female responsiveness (referred to as 'female unpaired response') was scored on a five-point scale (following Forstmeier 2007) (21): where -1 represents a clear rejection (involving aggression, threat, or fleeing) and +1 a clear acceptance (involving copulation solicitation, beak wiping, and ritualized hopping) with intermediate scores (-0.5, 0, +0.5) given for weaker or mixed responses (Fortmeier 2007; Forstmeier et al. 2011) (21, 23). For this study we combined 3,776 trials from the initial study (Forstmeier et al. 2011) (23) and 3,014 trials on selection line birds (see below), resulting in a total of 6,786 measures of 'male courtship rate' (four encounters with missing data were excluded) and 5,039 measures of 'female unpaired response' (74% of all trails; responsiveness could not be scored in 1,751 trials, typically when there is no male display). The trials involving 1,556 bachelor males and 1,441 bachelor females were carried out between July 2002 and December 2013. In these trials, males encountered on average 4.36 ± 1.3 SD (range 2-8) different females, and females encountered on average 4.54 ± 2.2 SD (range 1-14) different males (Table S1).

Selection on Male Courtship Rate

To verify previous findings (23) (Forstmeier et al. 2011), we established selection lines that were selected for divergent breeding values for male courtship rate starting in 2009 (some details see 36, 38) (Mathot et al. 2014, Wang et al. 2017b). We expected that, given the high genetic correlation between male courtship rate and female extra-pair mating behavior, the level of female promiscuity will change between lines as they diverge in male courtship rate (23) (Forstmeier et al. 2011).

Founder generation 'S0'

Before initiating the breeding of selection lines, we had measured the courtship rate of 585 males from four consecutive generations (P to F3, not including F4 birds; see (23) (Forstmeier et al. 2011)) in 2,922 trials. Using these measurements, we estimated breeding values for male courtship rate with a pedigree-based animal model. Breeding values of all individuals in the pedigree (n = 1219 from P to F3, including females) were calculated using VCE 6.0.2 (39) (Groeneveld et al. 2008). The single-trait permanent-environment animal-model was set up as follows: (1) 'Male courtship rate' was squared-root transformed to approach normality and used as the response variable (Table S1); (2) fixed effects were: male test day (four levels, from day one to day four), time of day (continuous, from 8:51 AM to 18:19 PM), the male inbreeding coefficient F (continuous, from 0 to 0.25) and the rearing environment of the male (two levels, either mixed-sex or unisex); (3) as random effects we included 'animal' (additive genetic effect),

'Male ID' (permanent environment effect, 585 levels), 'mother ID' (maternal effect, 203 levels), 'Test Batch ID' (periods of testing, 8 levels), and 'cohort ID' (periods of breeding, 6 levels).

Six breeding lines (two control, two high and two low lines) were started by choosing founder individuals with known breeding values for courtship rate (see above) from the pool that were still alive in May 2009 (n = 773; see Table S18). For each line, we set up 15 pairs to breed in one of the 90 randomly assigned cages (60×40 cm and 45 cm high) that were distributed over two breeding rooms (45 cages each). In case of individuals dying during the breeding, we also kept up to six replacement birds of each sex. In this founder generation (generation referred to as 'S0'), birds for the two control lines were chosen randomly from the entire pool before choosing the birds for the high and low lines. For the two 'high lines', we first selected 30 birds of each sex with the highest breeding values, and randomly allocated half of them to each replicate line. After that, we picked replacement birds of each sex with the next highest breeding values and distributed them randomly among the two lines. The two 'low lines' were set up in the same way, using the birds with the lowest breeding values.

Within each line, the allocation of the 30 individuals to form 15 breeding pairs was done in such a way as to minimize the level of inbreeding (see Table S18). Breeding in individual cages consisted of two rounds, together lasting about 14 months (from pair formation to independence of the last offspring). The pairs in each breeding round were allowed to breed until we obtained about 50 juveniles from each line. The partners of birds within each line were then swapped between the two breeding rounds (breeding cages again randomly assigned), in order to create maternal and paternal half sibs, thereby facilitating the separation of maternal effects from additive genetic effects. Juveniles of one breeding round went on to grow up (from 35 days of age to about 120 days of age) in one of two large mixed-sex peer groups, depending on the breeding room of origin (each containing 45 pairs from all lines). Across both rounds of breeding, the roughly 600 offspring were hence raised in one of four mixed-sex peer groups, each comprising about 75 males and 75 females from all lines.

Breeding generations 'S1' to 'S3'

Birds of the 'S0' generation produced 568 offspring (referred to as the pool of 'S1' generation) of which 546 survived until we were ready to start breeding the next generation (see Table S18). Male courtship rate and female unpaired response of these offspring were measured four times per individual (age of testing see Table S18), and then these new measurements were added to update the animal model for the calculation of new breeding values for all individuals (n = 1,929). The same fixed and random effects were included in this updated animal model, yet this time including 4,362 measurements of courtship rate from 947 males.

Selection of 'S1' breeders (15 pairs plus five replacement birds of each sex in each line) was carried out as before (random for control lines and based on breeding values for high and low lines; available pool see Table S18). Breeding pairs were formed in a way as to minimize and standardize the average inbreeding coefficient (equal mean F for the six lines, see Table S18). Specifically, in the most inbred line (high 2), we minimized inbreeding as much as possible, and pairs in the five other lines were chosen to match the mean value for this line. The mean inbreeding coefficients of resulting offspring for each line are given in Table S18.

The following generations 'S2' and 'S3' were bred following the same principles. For summary statistics see Table S18.

b) Aviary Breeding Experiments of 'S3' Birds

The third generation of the six selection lines (referred to as 'S3') consisted of 343 female and 338 male offspring, most of which had been phenotyped as usual for 'male courtship rate' and 'female unpaired response' in the cage experiments (see Table S18). To also measure other phenotypes that are more directly linked to extra-pair mating, we used a subset of 219 females and 217 males (about equally representing the six lines) that participated in communal-aviary breeding experiments.

We set up the same 9 breeding aviaries equipped with cameras, that had been used in the initial study (Forstmeier et al. 2011), for a period of 17 months (January 2014 to May 2015). Breeding was organized as follows: we created four consecutive testing cohorts because only 9 (rather than 36) aviaries were available at one time, each comprising 54 males and 54 females that were randomly drawn from the available pool of birds in each line (9 males and 9 females from each line per cohort, 216 of each sex in total, plus a few replacements, see below). Each group was then distributed to the nine aviaries such that (1) all birds within an aviary were unfamiliar with each other and (2) each aviary contained one male and one female from each selection line (9 aviaries x 6 lines corresponding to 54 individuals of each sex). Yet this procedure was possible only for the first 25 out of 36 experimental aviaries after which we had to start filling up a shortage of 'low 1' females and males with replacements from 'low 2', and later also replacing 'high 2' males with 'high 1' males. Hence aviaries were always balanced for containing 2 males and 2 females from each line type, but overall the number of tested birds per line and sex varied from 25 to 47 (see Table S18). With this set up, birds were given a choice of 6 potential mates (usually one from each line), yet social pairing appeared random with regard to line, so this issue was not considered further. Birds were given 7 weeks of time, which was sufficient for most birds to lay three clutches (nest boxes were provided from day 1 to day 45). All eggs laid were replaced by plastic eggs as soon as found and collected for later parentage assignment.

Clutches consisting of plastic eggs were removed after 10 days of incubation to allow the female to lay the next clutch. On day 49, individuals were separated by sex into different rooms for a two-week period, after which we initiated an identical second round of breeding, but with a different set of potential partners and extra-pair males (by swapping the six males of one aviary to the next). This allowed us to quantify the repeatability of the traits we have measured with different partners, and more importantly, allowed us to disentangle effects of 'Female ID' from 'Male ID' and 'Pair ID'. For this second round of breeding it was, however, no longer possible to ensure that all birds were unfamiliar to all opposite-sex individuals (on average 25% were familiar due to the joint rearing in one of four large natal peer groups). Across the four consecutive testing groups, one male and three females died during the first breeding round and were replaced by an individual from the same line in the second round, leading to a total of 217 males and 219 females participating in the experiments.

During the 2 x 7 weeks of breeding, we observed all birds (fitted with randomly assigned colored leg bands for individual identification) for signs of social pair bonding. Observations lasted about 30 min (for the 9 aviaries) and were carried out about 120 times per breeding round. We recorded all instances of allopreening, sitting in body contact or close to each other, and visiting a nest-box together. The start of a pair bond was defined as the first evidence of 'exclusive' bonding by the female to one male (i.e. >50% of bonding behaviours directed to one male; minimum 8 observations on this female-male combination; see Wang et al. 2017b (40) for details).

Following the initial study (Forstmeier et al. 2011) (23), we used video cameras to monitor the birds' courtships continuously in each aviary. Given that courtships were most frequently observed in the early morning, we always analyzed the first hour of video of every day during the breeding period, plus another two hours per day (randomly selected for each day). Thus, we screened a total of 10,656 hours of video (3h x 49.33 days x 9 aviaries x 2 breeding rounds x 4 testing cohorts), watching at 8-fold speed for detection of courtships (equal numbers of hours randomly allocated to two observers D.W. and K.M.), and found a total of 33,003 courtships. Apart from 10,614 courtships involving socially unpaired females (not analyzed here), we observed 9,121 courtships of paired females by extra-pair males (involving 206 females) for scoring 'female extra-pair response' and 13,268 courtships by the social partner (involving 200 females) for scoring 'female within-pair response'. For each courtship, K.M. scored female responsiveness as in the initial study (Forstmeier 2011) (23): threat or aggression toward the male (−1), flying away (−0.5), mixed or ambiguous signs (0), courtship hopping and beak wiping (+0.5), and copulation solicitation (+1). For joint analyses of the data from the initial study (Forstmeier et al. 2011) (23) and the present data from the selection lines, we also incorporated

the data from the initial study which contained 3,958 scores of 'female extra-pair response' (from 141 females) and 4,601 scores of 'female within-pair response' (from 143 females; Table S1) (Forstmeier et al. 2011) (23).

Paternity Analysis: 'female EPP', 'male EPP, and 'male EPE'

In total, 4,041 eggs were collected during the aviary breeding experiments of the 'S3' generation and these were placed into an incubator for 4 days in order to obtain embryonic tissue for parentage analysis. However, from 685 eggs no DNA was obtained (14 tiny eggs without yolk, 24 broken eggs, 632 apparently infertile eggs and 15 cases of lost samples or DNA concentration too low). The remaining 3,356 eggs were unambiguously assigned to parents using 15 microsatellite markers (Wang et al. 2017a) (35), but four eggs were only assigned to their mother (due to parthenogenesis, mosaicism, or siring by sperm from the previous experimental round). To quantify the level of female extra-pair paternity ('female EPP') we focus on a subset of 2,951 eggs that were laid by females with a clear social pair bond (i.e. after pairing). Of these, 726 eggs (24.6%) were sired by a male other than the partner. To obtain a comparable measure of male extra-pair paternity ('male EPP') that reflects the proportion of reproduction that happens outside the pair bond, we focus on a subset of 3,067 eggs that were sired by males with a clear social pair bond (i.e. after social pairing). Of these, 851 eggs (27.7%) had been laid by females other than the partner. Note that this measure of the proportion of male reproduction outside the social bond also depends on the partner's fecundity and fidelity, so it was not used in the initial study (Forstmeier et al. 2011) (23), and we here give it only for descriptive purposes (in Figure 2). Instead, we focus our quantitative genetic analyses, like in the initial study, on a measure of male extra-pair siring success ('male EPE') which is just the count of the extra-pair eggs (namely 851 eggs) that a male managed to sire while being involved in a social pair bond.

For joint analyses of the data from the initial study (Forstmeier et al. 2011) (23) and the present data from the selection lines, we incorporated additional measures of 'female EPP' from 2,253 eggs laid by 149 females and measures of 'male EPE' from 152 males (Forstmeier et al. 2011) (23, Table S1).

Female Fecundity

The quantification of 'female fecundity' from the current aviary breeding experiment of the S3 generation has been described in detail for a study on male mating preferences (35) (Wang et al. 2017a). In brief, 'female fecundity' is simply the count of all eggs that were laid by a female within one breeding round (here 45 days), based on a combination of genetic assignment of

maternity (3,356 eggs) and social assignment of eggs without DNA sample based on observations of nest attendance (610 eggs). For eggs with DNA sample, the social assignment proved to be correct in 93.1% of cases (false assignments resulted from egg dumping or nest take-over; Wang et al. 2017a), hence assignment errors appear negligible compared to the omission of all eggs without DNA sample. This resulted in 432 estimates of female fecundity (216 females x 2 breeding rounds, yet involving 219 individuals) based on 3,966 assigned eggs (mean ± SD = 9.2 ± 5.1, range 0-22). In order to increase the statistical power for quantifying genetic covariance between female fecundity and measures of promiscuity, we also included very similar data on female fecundity from a total of seven other aviary breeding experiments with genetic parentage assignment that had been carried out between 2005 and 2017 (involving 6 generations), the first four of which had been summarized in the initial study of (Forstmeier et al. 2011) (23). This resulted in a total of 854 fecundity estimates for 461 individual females based on the assignment of 9,127 eggs (mean ± SD = 10.7 ± 6.8, range 0-38). Differences between the eight experiments were accounted for in the statistical analyses (see below).

Data Analysis

Sample sizes and descriptive statistics of the data used for quantitative genetic analyses are given in Table S1 (including the data from the initial study, 23, Forstmeier et al. 2011). In general, the present analyses follow closely those used in the initial study, except where we felt that an important random effect had been missed (e.g. 'Pair ID' and 'Clutch ID') or a fixed effect could be better modelled as random (e.g. 'Test Batch ID'). To examine whether conclusions of the initial study were dependent on such arbitrary decisions about model structure, we repeated the initial models with updated model structure.

a) Mixed-effect Models Testing Extra-pair Paternity Levels of the Selection Lines

To test whether the birds from high lines indeed had higher levels of extra-pair paternity than birds from low lines after three generations of selection on male courtship rate, we analysed individual levels of 'male EPP' (n = 188 males) and of 'female EPP' (n = 190 females). We used mixed-effect models in the lme4 package in R 3.4.0 (41, 42) (Bates et al. 2015; R Core Team 2015) to test for differences in EPP levels across the six selection lines. For each sex, the counts of extra-pair and within-pair eggs of an individual within each round were analyzed as the dependent variable (binomial model of counts using the 'cbind' function in R). As the fixed effect of interest, we fitted the 'selection regime' as a covariate with one degree of freedom (low lines = -1, control lines = 0, and high lines = 1). As random effects we fitted either 'Female ID' (for female EPP) or 'Male ID' (for male EPP), and always 'Selection Line ID' (six levels) as well

as 'Individual within breeding round ID' (each line in the data sheet as a separate level in order to control for overdispersion of counts within an individual's breeding round).

b) Statistical Approach to Fixed Effects for Quantitative Genetic Models

First, we used generalized linear mixed-effect models (Bates et al. 2015; R Core Team 2015) (41, 42) to investigate how each of the traits measured in this study depended on a range of fixed effects. Specific details given below refer to the joint data set (initial study plus data from selection lines).

'Male courtship rate'

For 'male courtship rate' (square-root transformed to approach normality, Table S1), we used a mixed-effect model with 'Male ID' and 'Test Batch ID' (19 levels) as random effects. These two random effects explained 46% and 13% of variance after accounting for fixed effects, respectively. 'Male courtship rate' declined significantly over consecutive test days, declined with time of day, declined with male inbreeding coefficient, and was higher for males from a mixed-sex rearing environment compared with the unisex (Table S2).

'Male EPE'

The number of extra-pair eggs that paired males sired within each breeding round ('Male EPE') was square-root transformed to approach normality, and was modelled as the dependent variable (Table S1). 'Male ID' and 'breeding year' (six levels) were included as random effects which explained 21% and 8% of the variation, respectively. 'Male EPE' increased strongly with the number of days that males have been paired. This fixed effect controls for variation in the duration of experiments and for periods where males are unpaired. 'Male EPE' also declined with male inbreeding coefficient (Table S2).

'Female unpaired response'

The mixed-effect model for 'female unpaired response' included 'Female ID' and 'Test Batch ID' (19 levels) as random effects, accounting for 37% and 13% of the variation, respectively. The responsiveness of unpaired females to unfamiliar males differed significantly over consecutive test days: using the first day as reference level, the female responsiveness declined significantly in the second testing day, yet, increased significantly in the third testing day and showed no difference in the fourth day. Further, the responsiveness was higher in females that were reared in mixed-sex as opposed to unisex (Table S2).

'Female extra-pair response'

In the joint data sets of 'female extra-pair response', females interacted with an average of 5.54 ± 2.42 (range 1-12; 97% of 346 females with two or more) different extra-pair males. The model included three random effects: 'Female ID' (accounting for 5% of variance), 'Pair ID' (i.e. the combination of identities of the courted female and the couting extra-pair male; 23% of variance) and 'year' (1% of variance). The 'female extra-pair response' declined strongly with the time after dawn and with the duration of the pair bond (days paired). Based on the initial study (Forstmeier et al. 2011), we presumed that 'female extra-pair response' varies over the fertile cycle with highest responsiveness at 3 days before the start of egg laying (day 0) and with a continuous decline over the laying sequence. Hence, the fertile cycle was again modeled as the number of days away from day -3 (> = 5 coded as 5, 6 levels: from 0 to 5), and the laying sequence was modeled as the number of eggs laid in the previous 5 days. Although all courtships since 2007 had been scored by the same observer (K.M.), we had data from two additional observers in 2006, so we included observer ID as a fixed effect, showing that the scores of female extra-pair response varied slightly among the three observers (Table S2).

'Female within-pair response'

The 'female within-pair response' was modeled the same as 'female extra-pair response'. The three random effects 'Female ID', 'Pair ID', and 'year' accounted for 1%, 15% and 5% of variance, respectively. The 'female within-pair response' declined strongly with the time after dawn but increased strongly with the duration of the pair bond (days paired). Within-pair responsiveness varied over the fertile cycle just as extra-pair responsiveness did (Table S2).

'Female EPP'

'Female EPP' (for each egg laid by a paired female, modeled as 0 = within-pair and 1 = extra-pair, 5,194 eggs in total) was modeled as a binomial dependent variable in a generalized linear mixed-effect model. We included the random effects 'Female ID', 'Pair ID' (i.e. the combination of identities of the social partners) and 'clutch ID' (a clutch was defined as having no laying gaps longer than 4 days). We found that the variance components of 'Pair ID' (15.1) and 'clutch ID' (116.7) were considerably larger than what was explained by 'Female ID' (3.5×10^{-15}). As fixed effects, we included: (1) sex ratio in the aviary (3 levels; only relevant for data from 2005 and 2006), (2) the inbreeding coefficient of the female's social partner, (3) the number of days that the female had been paired (i.e. pair bond duration up to the date of egg laying). 'Female EPP' decreased with the duration of the pair bond, was higher when the sex-ratio was female-biased but was not influenced by the male partner's inbreeding coefficient (Table S2).

'Female fecundity'

'Female fecundity' (number of eggs laid per breeding round) was square-root transformed to approach normality, and modelled with the random effects of 'Female ID' (explaining 45% of the variance) and 'experiment ID' (18 levels after differentiating testing cohorts and breeding rounds; 10% of variance). 'Female fecundity' further increased with the number of days that females were present in an experiment (mean ± SD = 60 ± 23 days, range 1–112), and decreased with female age (mean ± SD = 735 ± 285 days, range 265–1511 days; Table S2).

c) Quantitative Genetic Analysis

We used animal models to carry out quantitative genetic analyses, closely following the initial study (23) (Forstmeier et al. 2011). For greater reliability we implemented both a restricted maximum likelihood (REML) and a Bayesian approach using Monte Carlo-Markov Chain (MCMC) to calculate the parameters. Likelihood-based animal models were carried out using VCE 6.0.2 (Groeneveld et al. 2008) (43) and Bayesian-based animal models were carried out with the package MCMCglmm in R 3.4.0 (Hadfield 2010) (44). Within each type of model (VCE or MCMCglmm), we varied the units of analysis (raw data representing single observations *vs.* individual mean trait estimates based on BLUPs, i.e. best linear unbiased predictions).

Specifically, to test the 'intersexual pleiotropy' hypothesis, we used four versions of animal models (like in Forstmeier et al. 2011) (23) to estimate the heritability and genetic correlations between aspects of male and female extra-pair mating behavior (five traits: 'male courtship rate', 'male EPE', 'female unpaired response', 'female extra-pair response' and 'female EPP'). These four versions of animal models were: a five-trait permanent-environment model (i.e. with repeated measures on individuals) in VCE (model I); a five-trait permanent-environment model in MCMCglmm (model II); a five-trait model on individual estimates in VCE (model III); and a five-trait model on individual estimates in MCMCglmm (model IV). For models III and IV, individual estimates were BLUPs that were extracted from the mixed-effect models shown in Table S2 (see details in Statistical Approach to Fixed Effects). The above models I to IV were based on the joint data from the initial study (Forstmeier, et al. 2011) (23) plus the follow-up data from the selection lines. For comparison between earlier and new findings we also ran models I and II on the respective subsets of data (initial data: presented as models V and VI which are updated for model structure compared to the ones published previously; new data: models VII and VIII).

To test the 'intrasexual pleiotropy' hypothesis, we used another four versions of animal models (similar to models I to IV above) to estimate the heritability and genetic correlations within the female sex (five traits: 'female fecundity', 'female unpaired response', 'female extra-pair response', 'female within-pair response' and 'female EPP'). These four versions of animal model

were: a five-trait permanent-environment model in VCE (model IX); a five-trait permanent-environment model in MCMCglmm (model X); a five-trait model on individual estimates in VCE (model XI); and a five-trait model on individual estimates in MCMCglmm (model XII). These models were all based on the joint data (initial plus new).

References

1. Petrie M & Kempenaers B (1998) Extra-pair paternity in birds: explaining variation between species and populations. *Trends Ecol Evol* 13(2):52-58.
2. Griffith SC, Owens IPF, & Thuman KA (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Mol Ecol* 11(11):2195-2212.
3. Westneat DF & Stewart IRK (2003) Extra-pair paternity in birds: Causes, correlates, and conflict. *Annu Rev Ecol Evol S* 34:365-396.
4. Forstmeier W, Nakagawa S, Griffith SC, & Kempenaers B (2014) Female extra-pair mating: adaptation or genetic constraint? *Trends Ecol Evol* 29(8):456-464.
5. Maldonado‐Chaparro AA, Montiglio PO, Forstmeier W, Kempenaers B, & Farine DR (2018) Linking the fine‐scale social environment to mating decisions: a future direction for the study of extra‐pair paternity. *Biological Reviews*.
6. Albrecht T*, et al.* (2007) Extrapair paternity and the opportunity for sexual selection in long-distant migratory passerines. *Behav Ecol* 18(2):477-486.
7. Webster MS, Tarvin KA, Tuttle EM, & Pruett-Jones S (2007) Promiscuity drives sexual selection in a socially monogamous bird. *Evolution* 61(9):2205-2211.
8. Sheldon BC & Mangel M (2014) BEHAVIOURAL ECOLOGY Love thy neighbour. *Nature* 512(7515):381-382.
9. Hsu YH, Schroeder J, Winney I, Burke T, & Nakagawa S (2015) Are extra‐pair males different from cuckolded males? A case study and a meta‐analytic examination. *Mol Ecol* 24(7):1558-1571.
10. Fox CW & Rauter CM (2003) Bet-hedging and the evolution of multiple mating. *Evol Ecol Res* 5(2):273-286.
11. Szulkin M, Stopher KV, Pemberton JM, & Reid JM (2013) Inbreeding avoidance, tolerance, or preference in animals? *Trends Ecol Evol* 28(4):205-211.
12. Kempenaers B (2007) Mate choice and genetic quality: A review of the heterozygosity theory. *Adv Stud Behav* 37:189-278.
13. Heg D, Ens BJ, Burke T, Jenkins L, & Kruijt JP (1993) Why Does the Typically Monogamous Oystercatcher (Haematopus-Ostralegus) Engage in Extra-Pair Copulations. *Behaviour* 126:247-289.
14. Lombardo MP & Thorpe PA (2000) Microbes in tree swallow semen. *J Wildlife Dis* 36(3):460-468.
15. Schmoll T, Schurr FM, Winkel W, Epplen JT, & Lubjuhn T (2009) Lifespan, lifetime reproductive performance and paternity loss of within-pair and extra-pair offspring in the coal tit Periparus ater. *P Roy Soc B-Biol Sci* 276(1655):337-345.
16. Sardell RJ, Arcese P, Keller LF, & Reid JM (2012) Are There Indirect Fitness Benefits of Female Extra-Pair Reproduction? Lifetime Reproductive Success of Within-Pair and Extra-Pair Offspring. *Am Nat* 179(6):779-793.
17. Hsu YH, Schroeder J, Winney I, Burke T, & Nakagawa S (2014) Costly Infidelity: Low Lifetime Fitness of Extra-Pair Offspring in a Passerine Bird. *Evolution* 68(10):2873-2884.
18. Halliday T & Arnold SJ (1987) Multiple Mating by Females - a Perspective from Quantitative Genetics. *Anim Behav* 35:939-941.
19. Kirkpatrick M & Barton NH (1997) The strength of indirect selection on female mating preferences. *P Natl Acad Sci USA* 94(4):1282-1286.
20. Arnqvist G & Kirkpatrick M (2005) The evolution of infidelity in socially monogamous passerines: The strength of direct and indirect selection on extrapair copulation behavior in females. *Am Nat* 165(5):S26-S37.

21. Forstmeier W (2007) Do Individual Females Differ Intrinsically in Their Propensity to Engage in Extra-Pair Copulations? *Plos One* 2(9).
22. Bolund E, Schielzeth H, & Forstmeier W (2012) Singing activity stimulates partner reproductive investment rather than increasing paternity success in zebra finches. *Behav Ecol Sociobiol* 66(6):975-984.
23. Forstmeier W, Martin K, Bolund E, Schielzeth H, & Kempenaers B (2011) Female extrapair mating behavior can evolve via indirect selection on males. *P Natl Acad Sci USA* 108(26):10608-10613.
24. Reid JM, Arcese P, Sardell RJ, & Keller LF (2011) Heritability of female extra-pair paternity rate in song sparrows (Melospiza melodia). *P Roy Soc B-Biol Sci* 278(1708):1114-1120.
25. Wilson A & Poissant J (2016) Quantitative genetics in natural populations. *The encyclopedia of evolutionary biology* 3:361-371.
26. Reid JM (2012) Predicting evolutionary responses to selection on polyandry in the wild: additive genetic covariances with female extra-pair reproduction. *P Roy Soc B-Biol Sci* 279(1747):4652-4660.
27. Reid JM, Arcese P, Keller LF, & Losdat S (2014) Female and Male Genetic Effects on Offspring Paternity: Additive Genetic (Co)Variances in Female Extra-Pair Reproduction and Male Paternity Success in Song Sparrows (Melospiza Melodia). *Evolution* 68(8):2357-2370.
28. Cohen J (1988) Statistical power analysis for the behavioral sciences 2nd edn. (Erlbaum Associates, Hillsdale).
29. Sgro CM, Chapman T, & Partridge L (1998) Sex-specific selection on time to remate in Drosophila melanogaster. *Anim Behav* 56:1267-1278.
30. Eady PE, Wilson N, & Jackson M (2000) Copulating with multiple mates enhances female fecundity but not egg-to-adult survival in the bruchid beetle Callosobruchus maculatus. *Evolution* 54(6):2161-2165.
31. Swallow JG, Carter PA, & Garland T (1998) Artificial selection for increased wheel-running behavior in house mice. *Behav Genet* 28(3):227-237.
32. Lynch M & Walsh B (1998) *Genetics and analysis of quantitative traits* (Sinauer Sunderland, MA).
33. Dietrich V, Schmoll T, Winkel W, Epplen JT, & Lubjuhn T (2004) Pair identity - An important factor concerning variation in extra-pair paternity in the coal tit (Parus ater). *Behaviour* 141:817-835.
34. Ihle M, Kempenaers B, & Forstmeier W (2015) Fitness Benefits of Mate Choice for Compatibility in a Socially Monogamous Species. *Plos Biol* 13(9).
35. Wang D, Kempenaers N, Kempenaers B, & Forstmeier W (2017) Male zebra finches have limited ability to identify high-fecundity females. *Behav Ecol* 28(3):784-792.
36. Wang D, Forstmeier W, & Kempenaers B (2017) No mutual mate choice for quality in zebra finches: Time to question a widely held assumption. *Evolution* 71(11):2661-2676.
37. Forstmeier W, Segelbacher G, Mueller JC, & Kempenaers B (2007) Genetic variation and differentiation in captive and wild zebra finches (Taeniopygia guttata). *Mol Ecol* 16(19):4039-4050.
38. Mathot KJ, Martin K, Kempenaers B, & Forstmeier W (2013) Basal metabolic rate can evolve independently of morphological and behavioural traits. *Heredity* 111(3):175-181.
39. Groeneveld E, Kovac M, & Mielenz N (2008) VCE User's Guide and Reference Manual Version 6.0.; 2010. *Google Scholar*.
40. Wang DP, Forstmeier W, & Kempenaers B (2017) No mutual mate choice for quality in zebra finches: Time to question a widely held assumption. *Evolution* 71(11):2661-2676.
41. Bates D, Machler M, Bolker BM, & Walker SC (2015) Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 67(1):1-48.
42. R Core Team (2015) R: a language and environment for statistical computing. Vienna, Austria; 2015).

43. Groeneveld E (2010) VCE user's guide and reference manual version 6.0. *ftp://ftp. tzv. fal. de/pub/vce6/doc/vce6-manual-3.1-A4. pdf*.
44. Hadfield JD (2010) MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *J Stat Softw* 33(2):1-22.

Funding

This work was supported by the Max Planck Society (to BK) and the China Scholarship Council (CSC; stipend to DW).

Acknowledgements

We thank Melanie Schneider for molecular work, and Sonja Bauer, Andrea Kortner, Jane Didsbury, and Petra Neubauer for animal care.

Author contributions

W.F., D.W. and B.K. conceived the study. D.W., W.F. and K.M. collected the data. D.W. and W.F. analyzed the data. D.W. wrote the manuscript with input from W.F., B.K. and A.W..

Table 1. Estimates of genetic correlation between aspects of male and female extra-pair mating using data from the initial study (Forstmeier et al. 2011) or new data from selection lines. Shown are parameter estimates ± standard errors (SE). Note that we updated the model structure for the re-analysis of the initial data to also include the random effects of clutch and pair identity (see Methods and Supplement). Estimates from animal models in VCE (Groeneveld 2010) (43) are based on restricted maximum likelihood (Table S10, Table S12). Estimates from animal models in MCMCglmm (Hadfield 2010) (44) are based on a Bayesian approach using Monte Carlo-Markov Chain modelling (Table S11, Table S13). 'EPP' stands for the proportion of eggs of a female that are sired by extra-pair males, and 'EPE' stands for male success in siring extra-pair eggs.

Male trait	Female trait	VCE initial data	VCE new data	MCMCglmm initial data	MCMCglmm new data
Courtship rate	Extra-pair response	0.885 ± 0.083	0.012 ± 0.084	0.562 ± 0.172	0.000 ± 0.169
Courtship rate	EPP	0.765 ± 0.197	0.060 ± 0.090	0.424 ± 0.258	0.069 ± 0.192
EPE	Extra-pair response	0.872 ± 0.147	-0.508 ± 0.225	0.443 ± 0.272	-0.161 ± 0.335
EPE	EPP	0.941 ± 0.093	-0.536 ± 0.259	0.331 ± 0.300	-0.162 ± 0.342

Figure 1. Male courtship rate in selection lines over three successive generations (a) to (c). Here, the y-axis shows courtship rate of male offspring (square-root transformed seconds in a 5-min trial, averaged across 4 trials per male) over their parents' breeding value for male courtship rate (x-axis). Three types of selection lines (high, control, and low) are shown across three generations of offspring from 'S1' to 'S3'. The parents' breeding values were estimated prior to breeding (without information on offspring phenotypes) from a single-trait permanent environment animal model in VCE. Equations of ordinary least square regression lines are shown.

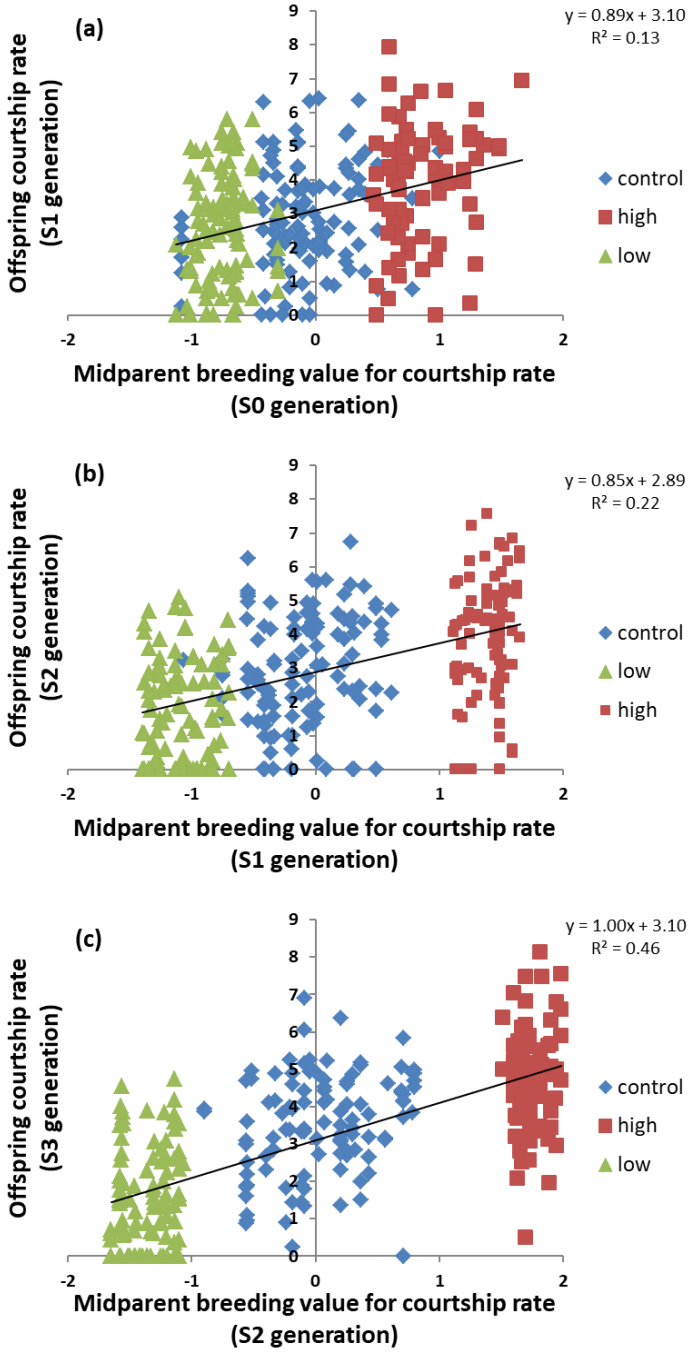


Figure 2. Weighted averages (± SE) of levels of extra-pair paternity by each of the six selection lines in aviary breeding experiments (data on ca. 3,000 eggs from the 'S3' generation). Here, male extra-pair paternity (x-axis) is the proportion of eggs that socially paired males sire outside their pair bond. Likewise, 'female extra-pair paternity (y-axis) is the proportion of eggs laid by socially paired females that are sired by males other than the partner.

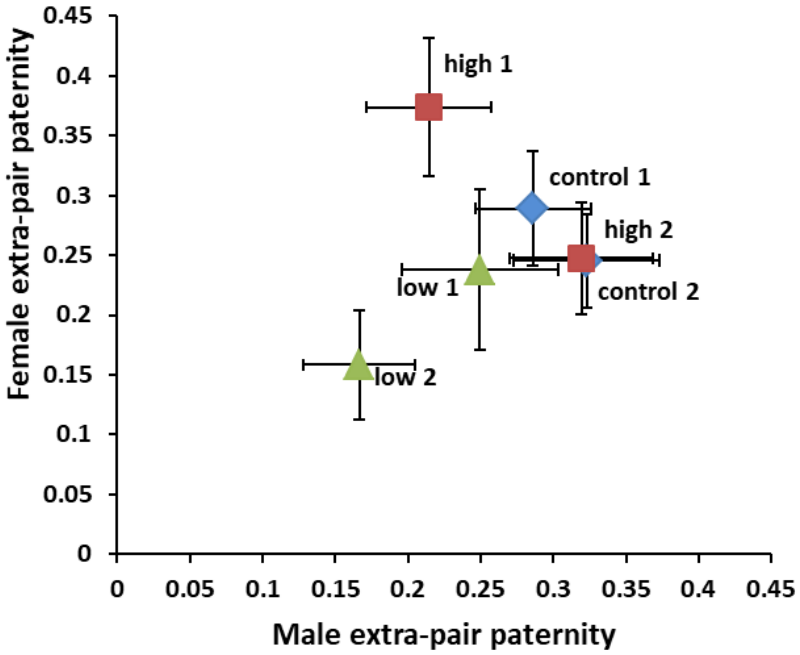


Figure 3. Repeatability of female extra-pair responsiveness (left panel) and female levels of extra-pair paternity (EPP, right panel) across two social pair bonds and social environments (1st vs 2nd round). Here, weighted (by the geometric mean of the two rounds) ordinary least square regression lines are shown (slope $\beta = 0.37 \pm 0.09$ and $\beta = 0.24 \pm 0.09$). Dot size refers to the geometric mean of the relevant sample sizes in the two breeding rounds (number of extra-pair courtships and number of eggs laid, respectively).

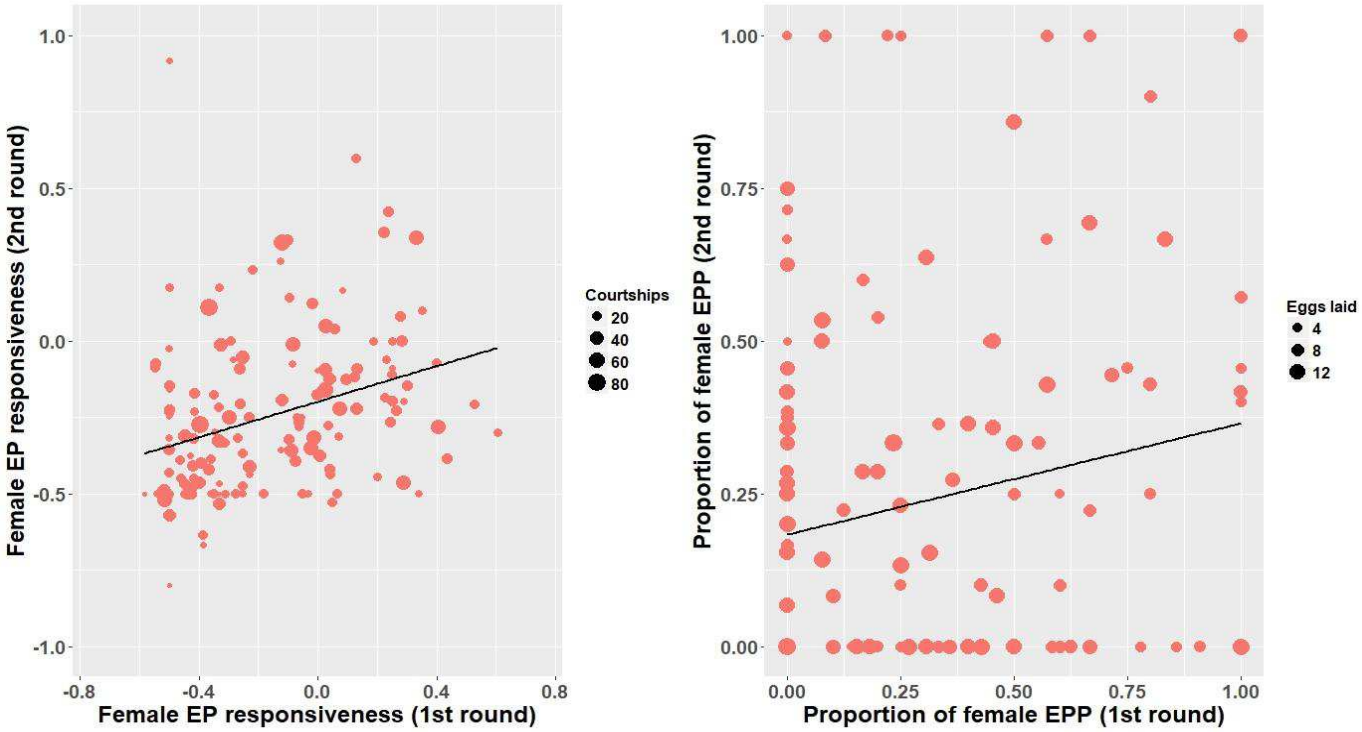


Figure 4. Female extra-pair mating behavior in relation to their estimated breeding value for male courtship rate. Panels (a) and (b) are based on data from the initial study, panels (c) and (d) are based on data from the selection lines ('S3' generation). (a) The average responsiveness of 141 females when courted by extra-pair males (total n = 3,958 courtships) in relation to their estimated breeding value for male courtship rate. Breeding values are from a single-trait permanent environment model conducted in VCE (based on courtship rates from 800 male relatives). Dot size refers to the number of extra-pair courtships ('Courtships') observed for each female (range: 1–138, median: 19). A weighted (by the number of 'Courtships') regression line is shown (slope $\beta = 0.14 \pm 0.03$). (b) The average proportion of extra-pair paternity (EPP) among the eggs laid by 149 females (total n = 2,253 eggs) in relation to their estimated breeding value for male courtship rate. Dot size refers to the number of eggs laid by each female (range: 1–45, median: 14). A weighted (by the number of eggs) regression line is shown ($\beta = 0.10 \pm 0.04$). (c) The average responsiveness of 205 females from selection lines (control, high, low) when courted by extra-pair males (total n = 9,117 courtships) in relation to their estimated breeding value for male courtship rate. Breeding values are from a single-trait permanent environment model conducted in VCE (based on courtship rates from 1,651 male relatives). Dot size refers to the number of extra-pair courtships ('Courtships') observed for each female (range: 1–219, median: 33). A weighted regression line is shown ($\beta = 0.02 \pm 0.01$). (d) The average proportion of extra-pair paternity among the eggs laid by 190 females from selection lines (control, high, low; n = 2,951 eggs) in relation to their estimated breeding value for male courtship rate. Dot size refers to the number of eggs laid by each female (range: 1–32, median: 15). A weighted regression line is shown ($\beta = 0.03 \pm 0.01$).

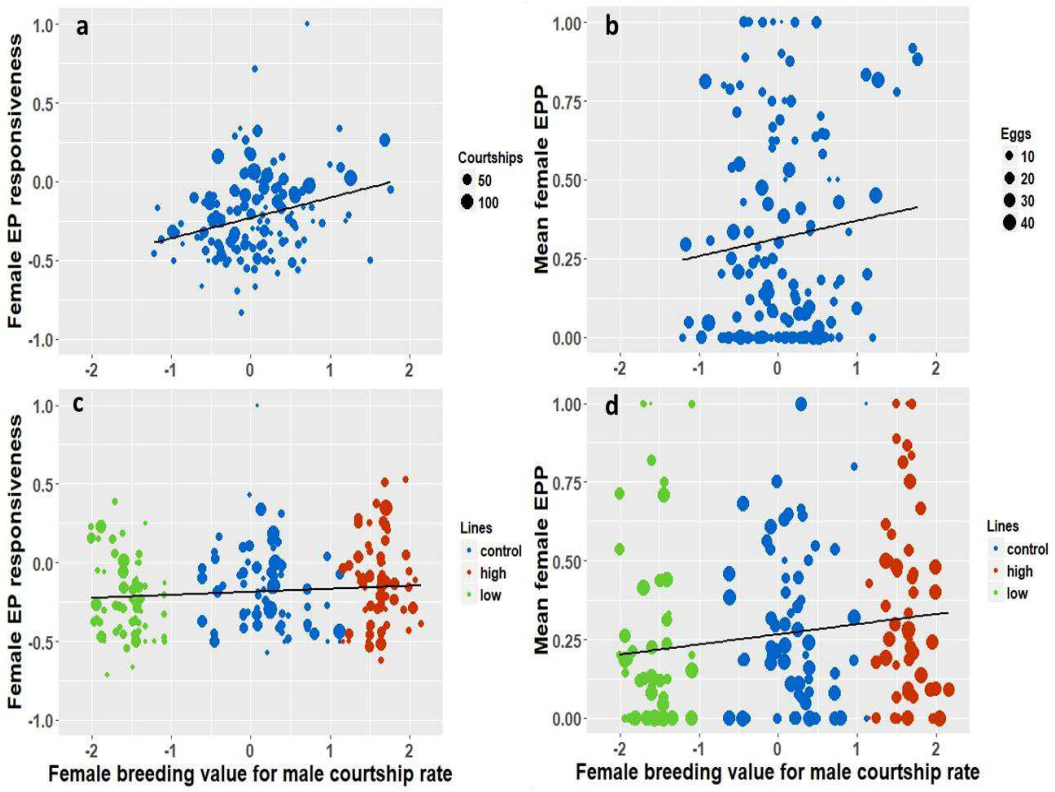
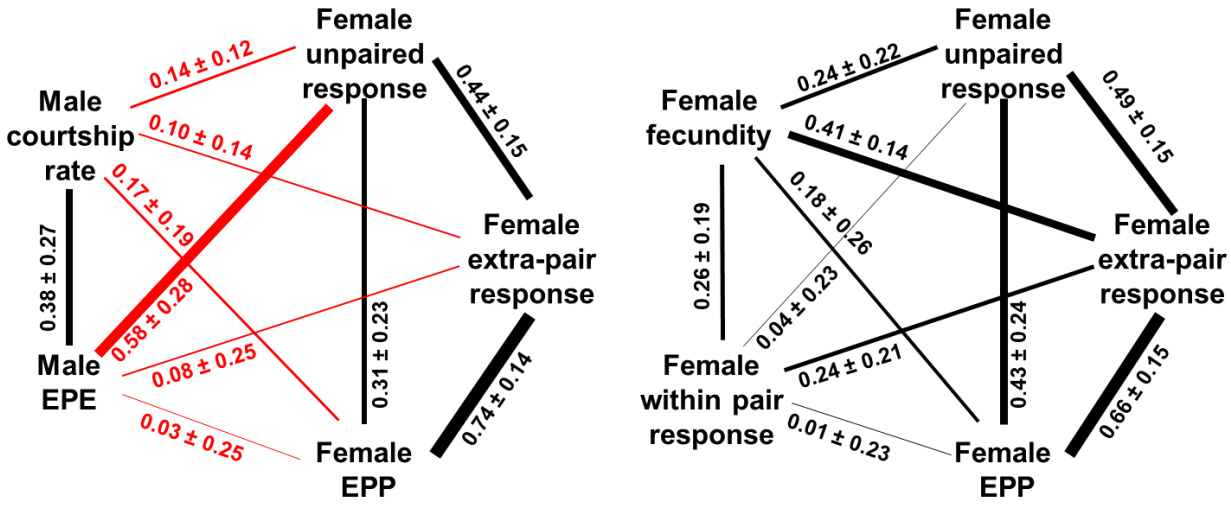

a
b
c
d
Female EP responsiveness
Mean female EPP
Female breeding value for male courtship rate
Courtships
50
100
Eggs
10
20
30
40
Lines
control
high
low

Figure 5. Estimates of genetic correlation between aspects of male and female extra-pair mating behavior (left panel) and among female traits (right panel). Median estimates of genetic correlations (± SE) are from four versions of animal models (Table S6 to S9, left panel; Table S14 to S17, right panel) and are based on the joint data from the initial study plus the data from selection lines. Between-sex genetic correlations are shown in red and within-sex genetic correlations are shown in black. The thickness of lines reflects the strength of correlation. EPP = extra-pair paternity; EPE = success in siring extra-pair eggs.



Genetic constraints of female promiscuity: male corollary or independent trajectory?

Daiping Wang, Katrin Martin, Wolfgang Forstmeier, Bart Kempenaers

SUPPLEMENT

Table S1. Descriptive statistics of traits used in quantitative genetic analyses. Sample sizes and distributions of the original measurements (Raw data), the transformed values (square-root transformation for 'male courtship rate', 'male EPE', and female fecundity, and logit transformation for 'female EPP'), and of male or female random effect estimates (BLUPs) extracted from the mixed models (shown in Tables S2). For BLUPs the percentage of variance explained by the random effect (male or female identity) is given.

		Male courtship rate	Male EPE	Female unpaired response	Female extra-pair response	Female within-pair response	Female EPP	Female fecundity
N	Individuals	1,556	369	1,441	346	343	339	461
	Measurements	6,786	634[a]	5,039	13,075	17,867	5,194[b]	854[a]
Raw data	Mean	19.1 [c]	2.71 [d]	-0.22	-0.15	0.46	0.27	10.69 [e]
	SD	18.8 [c]	5.00[d]	0.61	0.56	0.63	0.44	6.81[e]
	# of levels	112	29	5	5	5	2	33
	Range	0-144 [c]	0-40[d]	-1 to +1	-1 to +1	-1 to +1	0, 1	0-38[e]
Transformed	Transformation	sqrt	sqrt	-	-	-	logit	sqrt
	Mean	3.61	1.02	-0.22	-0.15	0.46		2.96
	SD	2.47	1.29	0.61	0.56	0.63		1.38
	Range	0-12	0-6.3	-1 to +1	-1 to +1	-1 to +1		0-6.16
	Assumed distribution	Normal	Normal	Normal	Normal	Normal	Binomial	Normal
BLUPs	Mean	0	0	0	0	0	0	0
	SD	1.51	0.32	0.31	0.08	0.02	0	0.69
	Range	-3.93 to 5.39	-0.63 to 1.16	-0.84 to 0.97	-0.15 to 0.32	-0.07 to 0.05	-10^{-16} to 10^{-15}	-2.19 to 1.26
	% phenotypic variance [f]	46.0%	20.6%	37.3%	5.3%	1.0%	0%	44.9%

(a) Number of male or female breeding rounds
(b) Number of eggs
(c) In seconds
(d) Number of extra-pair eggs sired per male breeding round
(e) Number of eggs laid
(f) Of transformed values after controlling for fixed effects

Table S2. Estimates of fixed effects on traits used in quantitative genetic analyses. Parameter estimates for fixed effects on two male and five female traits related to extra-pair mating. For continuous predictors (covariates, marked by "(C)") we give slope estimates in relation to the units of change in the predictor. For factors we give estimates for each level relative to the first level (reference). Parameter estimates and standard errors (SE) are from seven univariate mixed-effect models performed in R (LMER estimate). For comparison, parameter estimates from five-trait permanent-environment animal models are given, one performed in VCE (Model I) and one in MCMCglmm (Model II).

Dependent trait	Level or Covariate	Reference level or scaling of the covariate	LMER estimate	SE	VCE estimate	MCMC estimate
Male courtship rate	Intercept		4.820	0.325	4.872	3.538
	Test day 2	Test day 1	-0.496	0.040	-0.496	-0.496
	Test day 3	Test day 1	-1.213	0.111	-1.215	-1.207
	Test day 4	Test day 1	-1.403	0.093	-1.408	-1.399
	Mixed-sex rearing [(a)]	Uni-sex rearing	0.424	0.217	0.588	0.599
	Male F [(b)] (C)	Per 0.25F	-0.988	0.247	-1.223	-1.190
	Daytime (C)	Per 1h	-0.067	0.014	-0.070	-0.069
Male EPP	Intercept		-0.117	0.248		-0.047
	Days paired (C)	Per 100 days	2.022	0.253	1.838	1.946
	Male F [(b)] (C)	Per 0.25 F	-0.542	0.218	-0.769	-0.556
Female unpaired response	Intercept		-0.288	0.069	-0.288	-0.290
	Test day 2	Test day 1	-0.087	0.016	-0.088	-0.088
	Test day 3	Test day 1	0.115	0.020	0.116	0.115
	Test day 4	Test day 1	-0.003	0.020	-0.002	-0.003
	Mixed-sex rearing [(a)]	Uni-sex rearing	0.165	0.051	0.225	0.208
Female extra-pair response	Intercept		0.126	0.049	0.130	0.106
	Author KM [(c)]	Author EB	-0.127	0.036	-0.121	-0.124
	Author WF	Author EB	-0.083	0.039	-0.081	-0.081
	Log time (C)	Per 9 min [(d)]	-0.056	0.005	-0.053	-0.055
	Log days paired (C)	Per 9 days [(e)]	-0.041	0.012	-0.039	-0.037
	Days from day -3 (C)[(f)]	Per 1 day	-0.027	0.003	-0.026	-0.027
	Eggs in last 5 days (C)	Per 1 egg	-0.055	0.005	-0.057	-0.056
Female within-pair response	Intercept		0.126	0.049	0.054	0.111
	Author KM [(c)]	Author EB	-0.127	0.036	-0.122	-0.125
	Author WF	Author EB	-0.083	0.039	-0.082	-0.083
	Log time (C)	Per 9 min [(d)]	-0.056	0.005	-0.052	-0.056

	Log days paired (C)	Per 9 days [e]	-0.041	0.012	-0.042	-0.039
	Days from day-3 (C)[f]	Per 1 day	-0.027	0.003	-0.026	-0.027
	Eggs in last 5 days (C)	Per 1 egg	-0.055	0.005	-0.058	-0.055
Female EPP	Intercept		-6.246[g]	1.115	0.422	0.609
	Sex ratio 0.5	Sex ratio 0.4	-0.410[g]	0.774	-0.027	-0.041
	Sex ratio 0.6	Sex ratio 0.4	2.301[g]	1.052	0.183	0.177
	Log days paired (C)	Per 9 days [e]	-1.577[g]	0.433	-0.098	-0.103
	Partner F [b] (C)	Per 0.25 F	0.785[g]	0.811	0.101	0.097
Female fecundity	Intercept		1.426	0.365	1.024	1.470
	Sex ratio 0.5	Sex ratio 0.4	0.260	0.201	0.261	0.216
	Sex ratio 0.55	Sex ratio 0.4	0.148	0.342	0.185	0.099
	Sex ratio 0.6	Sex ratio 0.4	-0.049	0.215	-0.039	-0.079
	Age (C)[h]	Per year	-0.337	0.117	-0.361	-0.338
	Days present (C)[i]	Per day	0.032	0.003	0.038	0.032

(a) Birds were reared in either mixed-sex or uni-sex peer groups
(b) Inbreeding coefficient F, the parameter estimate is for a change in F of 0.25 units
(c) Observer scoring responsiveness
(d) Log(x+1) transformed time in minutes; the first unit of time has passed 9 min after lights on, the second after 99 min
(e) Log(x+1) transformed time paired in days; the first unit of time has passed 9 days after pair formation, the second after 99 days
(f) The number of days between the courtship and the day three days before the start of egg laying (with values ≥5 coded as 5)
(g) Parameter estimates are on the logit scale and hence not directly comparable to the VCE estimate
(h) Female age in years at start of experiment
(i) The number of days a female was present in a breeding experiment

Table S3. Direct response to selection. Average male courtship rate (± SE) and sample sizes for each selection line in the 'S3' generation (after three consecutive generations of selection).

Selection line	Courtship rate (square root transformed seconds)	Courtship rate (seconds)	Number of males
High 1	4.7 ± 0.19	22.1	53
High 2	4.7 ± 0.23	22.1	34
Control 1	3.6 ± 0.16	13.0	62
Control 2	3.4 ± 0.21	11.6	51
Low 1	1.4 ± 0.21	2.0	40
Low 2	1.6 ± 0.16	2.6	63

Table S4. Indirect response to selection for 'female EPP'. Results from a linear mixed-effect model testing for a difference in female extra-pair paternity levels across the six selection lines in the S3 generation. The dependent variable is the relative numbers of extra-pair versus within-pair eggs (using the 'cbind' function in R) within each female breeding round (n = 325). The fixed effect 'Selection regime' was modeled with one degree of freedom (coding low lines as -1, control lines as 0, and high lines as +1). The random effect 'Female ID' reflects the individual repeatability across different partners and social environments (variance component 'Var'), while the random effect of selection-line identity ('Line ID') is meant to control for non-independence of data within lines. The random effect of female breeding rounds ('Female rounds') controls for overdispersion in the binomial counts. The negative intercept (on the logit scale) reflects that extra-pair paternity levels are below 50%. The model suggests a significant effect of 'Selection regime' on levels of extra-pair paternity.

Variable	Estimate Var or β ± SE	z	*p*
Random effects			
Female ID (190 levels)	1.546		
Line ID (6 levels)	0.000		
Female rounds (325 levels)	4.470		
Fixed effects			
Intercept	-2.039 ± 0.19	-10.5	10^{-16}
Selection regime (1df)	0.698 ± 0.23	3.10	0.002

Table S5. Indirect response to selection for 'male EPP'. Results from a linear mixed-effect model testing for a difference in male extra-pair paternity levels across the six selection lines in the S3 generation. The dependent variable is the relative numbers of extra-pair versus within-pair eggs (using the 'cbind' function in R) within each male breeding round (n = 319). The fixed effect 'Selection regime' was modeled with one degree of freedom (coding low lines as -1, control lines as 0, and high lines as +1). The random effect 'Male ID' reflects the individual repeatability across different partners and social environments (variance component 'Var'), while the random effect of selection-line identity ('Line ID') is meant to control for non-independence of data within lines. The random effect of male breeding rounds ('Male rounds') controls for overdispersion in the binomial counts. The negative intercept (on the logit scale) reflects that extra-pair paternity levels are below 50%. The model suggests a non-significant effect of 'Selection regime' on levels of extra-pair paternity.

Variable	Estimate Var or β ± SE	z	*p*
Random effects			
Male ID (188 levels)	0.676		
Line ID (6 levels)	0.000		
Male rounds (319 levels)	2.608		
Fixed effects			
Intercept	-1.751 ± 0.15	-12.1	10^{-16}
Selection regime (1df)	0.278 ± 0.17	1.68	0.09

Table S6. Animal model I, intersexual, initial + new data, VCE, raw data. Variance components and correlations estimated from a five-trait (two male and three female traits) permanent-environment animal model performed in VCE based on the joint data (initial and follow-up study). Variance components are standardized by the phenotypic variance (after controlling for fixed effects). Variance components ± SE are shown on the diagonal (heritabilities are highlighted in grey), correlations ± SE between pairs of traits are shown off the diagonal. Between-sex genetic correlations are highlighted in bold. The additive-genetic and permanent-environment components together reflect the individual repeatability of single units of observation. UP = unpaired (cage experiments), EP = extra-pair (aviary experiments), EPP = extra-pair paternity, EPE = extra-pair eggs. As for 'EP response', 'Pair ID' stands for combination of focal female and the courting extra-pair male. As for 'female EPP', 'Pair ID' stands for the social pair bond. Test Batch, year and clutch ID are also included as random effects. Parameters that do not apply to a trait or cannot be estimated are marked with "---".

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.198 ± 0.029	0.457 ± 0.324	**0.124 ± 0.063**	**0.067 ± 0.114**	**0.084 ± 0.133**
	Male EPE		0.025 ± 0.021	**0.744 ± 0.710**	**-0.002 ± 0.235**	**-0.062 ± 0.294**
	Female UP response			0.213 ± 0.028	0.512 ± 0.121	0.449 ± 0.162
	Female EP response				0.043 ± 0.019	0.997 ± 0.013
	Female EPP					0.057 ± 0.026
Test Batch ID	Male courtship	0.177 ± 0.022	---	0.779 ± 0.056	---	---
	Male EPE		---	---	---	---
	Female UP response			0.110 ± 0.015	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.217 ± 0.029	0.167 ± 0.223	---	---	---
	Male EPE		0.183 ± 0.045	---	---	---
	Female UP response			0.181 ± 0.024	0.264 ± 0.489	0.363 ± 0.348
	Female EP response				0.022 ± 0.018	0.994 ± 0.028
	Female EPP					0.074 ± 0.024
Year	Male courtship	---	---	---	---	---
	Male EPE		0.071 ± 0.031	---	0.793 ± 0.339	---
	Female UP response			---	---	---
	Female EP response				0.009 ± 0.006	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.368 ± 0.023
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.223 ± 0.018	---
	Female EPP					0.214 ± 0.024
Residual	Male courtship	0.407 ± 0.025	---	---	---	---
	Male EPE		0.721 ± 0.051	---	---	---
	Female UP response			0.495 ± 0.023	---	---
	Female EP response				0.702 ± 0.014	---
	Female EPP					0.287 ± 0.009

Table S7. Animal model II, intersexual, initial + new data, MCMCglmm, raw data. Like Table S6, but model performed using MCMCglmm.

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.196 ± 0.031	0.303 ± 0.237	**0.143 ± 0.123**	**0.080 ± 0.155**	**0.082 ± 0.177**
	Male EPE		0.037 ± 0.021	**0.422 ± 0.309**	**0.146 ± 0.266**	**0.123 ± 0.297**
	Female UP response			0.213 ± 0.033	0.386 ± 0.169	0.322 ± 0.225
	Female EP response				0.045 ± 0.011	0.547 ± 0.141
	Female EPP					0.089 ± 0.027
Test Batch ID	Male courtship	0.163 ± 0.052	---	0.711 ± 0.145	---	---
	Male EPE		---	---	---	---
	Female UP response			0.119 ± 0.040	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.225 ± 0.027	0.366 ± 0.294	---	---	---
	Male EPE		0.098 ± 0.075	---	---	---
	Female UP response			0.176 ± 0.029	0.254 ± 0.209	0.284 ± 0.232
	Female EP response				0.030 ± 0.008	0.534 ± 0.146
	Female EPP					0.089 ± 0.029
Year	Male courtship	---	---	---	---	---
	Male EPE		0.070 ± 0.062	---	0.145 ± 0.375	---
	Female UP response			---	---	---
	Female EP response				0.063 ± 0.045	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.360 ± 0.020
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.204 ± 0.014	---
	Female EPP					0.183 ± 0.027
Residual	Male courtship	0.416 ± 0.028	---	---	---	---
	Male EPE		0.795 ± 0.084	---	---	---
	Female UP response			0.491 ± 0.027	---	---
	Female EP response				0.658 ± 0.033	---
	Female EPP					0.280 ± 0.011

Table S8. Animal model III, intersexual, initial + new data, VCE, BLUPs. Like Table S6, but based on estimates of individual average phenotypes (using BLUPs). Here, the decomposition of variance is only into an additive genetic component, maternal effect component, and a residual component. Note that the heritability estimates (highlighted in grey) refer to the heritability of average phenotypes rather than the heritability of single measures (as is the case for models using raw data).

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.341 ± 0.036	0.505 ± 0.307	**0.185 ± 0.151**	**0.127 ± 0.292**	**0.251 ± 0.295**
	Male EPE		0.035 ± 0.040	**0.822 ± 0.247**	**0.093 ± 0.309**	**-0.109 ± 0.178**
	Female UP response			0.328 ± 0.048	0.503 ± 0.452	0.220 ± 0.442
	Female EP response				0.270 ± 0.122	0.940 ± 0.148
	Female EPP					0.097 ± 0.068
Maternal	Male courtship	0.057 ± 0.022	---	-0.707 ± 0.338	---	---
	Male EPE		---	---	---	---
	Female UP response			0.058 ± 0.028	---	---
	Female EP response				---	---
	Female EPP					---
Residual	Male courtship	0.602 ± 0.038	0.086 ± 0.040	---	---	---
	Male EPE		0.965 ± 0.058	---	---	---
	Female UP response			0.614 ± 0.047	0.052 ± 0.113	0.110 ± 0.083
	Female EP response				0.730 ± 0.122	0.224 ± 0.091
	Female EPP					0.903 ± 0.068

Table S9. Animal model IV, intersexual, initial + new data, MCMCglmm, BLUPs. Like Table S8, but model performed using MCMCglmm. Note that maternal effects were not estimated for consistency with the initial study (Forstmeier et al. 2011).

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.377 ± 0.051	0.264 ± 0.164	**0.144 ± 0.121**	**0.117 ± 0.129**	**0.810 ± 0.201**
	Male EPE		0.152 ± 0.048	**0.252 ± 0.225**	**0.067 ± 0.161**	**0.247 ± 0.211**
	Female UP response			0.399 ± 0.058	0.311 ± 0.110	0.295 ± 0.235
	Female EP response				0.543 ± 0.060	0.212 ± 0.153
	Female EPP					0.027 ± 0.021
Residual	Male courtship	0.623 ± 0.051				
	Male EPE		0.848 ± 0.048			
	Female UP response			0.601 ± 0.058		
	Female EP response				0.457 ± 0.060	
	Female EPP					0.973 ± 0.021

Table S10. Animal model V, intersexual, initial data only, VCE, raw data. Like Table S6, but based on the data of the initial study (Forstmeier et al. 2011) only. Note the updated model structure compared to the previously published version (including Clutch ID, Pair ID, and Test Batch ID as a random rather than fixed effect).

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.185 ± 0.026	0.601 ± 0.144	**0.144 ± 0.063**	**0.885 ± 0.083**	**0.765 ± 0.197**
	Male EPE		0.063 ± 0.033	**0.614 ± 0.247**	**0.872 ± 0.147**	**0.941 ± 0.093**
	Female UP response			0.212 ± 0.027	0.547 ± 0.125	0.338 ± 0.233
	Female EP response				0.058 ± 0.018	0.890 ± 0.210
	Female EPP					0.043 ± 0.025
Test Batch ID	Male courtship	0.180 ± 0.039	---	0.778 ± 0.097	---	---
	Male EPE		---	---	---	---
	Female UP response			0.117 ± 0.031	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.226 ± 0.024	0.132 ± 0.103	---	---	---
	Male EPE		0.096 ± 0.033	---	---	---
	Female UP response			0.178 ± 0.018	-0.137 ± 0.702	0.421 ± 0.164
	Female EP response				0.003 ± 0.006	0.839 ± 0.408
	Female EPP					0.088 ± 0.058
Year	Male courtship	---	---	---	---	---
	Male EPE		0.024 ± 0.017	---	0.778 ± 0.097	---
	Female UP response			---	---	---
	Female EP response				0.117 ± 0.031	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.340 ± 0.025
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.207 ± 0.020	---
	Female EPP					0.237 ± 0.032
Residual	Male courtship	0.409 ± 0.026	---	---	---	---
	Male EPE		0.818 ± 0.041	---	---	---
	Female UP response			0.482 ± 0.035	---	---
	Female EP response				0.715 ± 0.014	---
	Female EPP					0.292 ± 0.010

Table S11. Animal model VI, intersexual, initial data only, MCMCglmm, raw data. Like Table S10, but model performed using MCMCglmm.

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.193 ± 0.031	0.592 ± 0.383	**0.160 ± 0.116**	**0.562 ± 0.172**	**0.424 ± 0.258**
	Male EPE		0.061 ± 0.043	**0.349 ± 0.303**	**0.442 ± 0.272**	**0.331 ± 0.300**
	Female UP response			0.210 ± 0.033	0.371 ± 0.153	0.224 ± 0.262
	Female EP response				0.066 ± 0.019	0.353 ± 0.209
	Female EPP					0.115 ± 0.041
Test Batch ID	Male courtship	0.161 ± 0.053	---	0.700 ± 0.157	---	---
	Male EPE		---	---	---	---
	Female UP response			0.126 ± 0.045	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.229 ± 0.028	0.120 ± 0.573	---	---	---
	Male EPE		0.027 ± 0.024	---	---	---
	Female UP response			0.176 ± 0.029	0.111 ± 0.228	0.320 ± 0.270
	Female EP response				0.027 ± 0.009	0.180 ± 0.234
	Female EPP					0.104 ± 0.042
Year	Male courtship	---	---	---	---	---
	Male EPE		0.021 ± 0.033	---	0.074 ± 0.436	---
	Female UP response			---	---	---
	Female EP response				0.091 ± 0.080	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.326 ± 0.029
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.178 ± 0.022	---
	Female EPP					0.176 ± 0.047
Residual	Male courtship	0.417 ± 0.029	---	---	---	---
	Male EPE		0.891 ± 0.057	---	---	---
	Female UP response			0.487 ± 0.030	---	---
	Female EP response				0.638 ± 0.061	---
	Female EPP					0.279 ± 0.019

Table S12. Animal model VII, intersexual, new data only, VCE, raw data. Like Table S6, but based on the data on 'Male EPE', 'Female EP response', and 'Female EPP' are from the S3 generation of selection lines only. Note that the full data was used for the traits measured in cage experiments, namely 'Male courtship rate' and 'Female UP response'.

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.199 ± 0.025	0.346 ± 0.131	**0.115 ± 0.060**	**0.012 ± 0.084**	**0.060 ± 0.090**
	Male EPE		0.054 ± 0.042	**0.206 ± 0.056**	**-0.508 ± 0.225**	**-0.536 ± 0.259**
	Female UP response			0.209 ± 0.016	0.724 ± 0.196	0.686 ± 0.252
	Female EP response				0.049 ± 0.019	0.997 ± 0.017
	Female EPP					0.084 ± 0.036
Test Batch ID	Male courtship	0.176 ± 0.040	---	0.774 ± 0.116	---	---
	Male EPE		---	---	---	---
	Female UP response			0.116 ± 0.035	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.217 ± 0.024	0.176 ± 0.071	---	---	---
	Male EPE		0.139 ± 0.047	---	---	---
	Female UP response			0.183 ± 0.018	0.416 ± 0.350	0.218 ± 0.473
	Female EP response				0.025 ± 0.014	0.978 ± 0.060
	Female EPP					0.051 ± 0.024
Year	Male courtship	---	---	---	---	---
	Male EPE		0.130 ± 0.008	---	0.035 ± 0.013	---
	Female UP response			---	---	---
	Female EP response				0.118 ± 0.009	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.389 ± 0.025
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.196 ± 0.014	---
	Female EPP					0.195 ± 0.034
Residual	Male courtship	0.408 ± 0.026	---	---	---	---
	Male EPE		0.677 ± 0.057	---	---	---
	Female UP response			0.493 ± 0.020	---	---
	Female EP response				0.612 ± 0.013	---
	Female EPP					0.281 ± 0.012

Table S13. Animal model VIII, intersexual, new data only, MCMCglmm, raw data. Like Table S12, but model performed using MCMCglmm. Note that the full data was used for the traits measured in cage experiments, namely 'Male courtship rate' and 'Female UP response'

		Male		Female		
Effects		Courtship	EPE	UP response	EP response	EPP
Genetic	Male courtship	0.204 ± 0.033	0.262 ± 0.252	**0.130 ± 0.118**	**0.000 ± 0.169**	**0.069 ± 0.192**
	Male EPE		0.049 ± 0.038	**0.015 ± 0.467**	**-0.161 ± 0.335**	**-0.162 ± 0.342**
	Female UP response			0.209 ± 0.038	0.443 ± 0.233	0.364 ± 0.293
	Female EP response				0.045 ± 0.017	0.554 ± 0.156
	Female EPP					0.102 ± 0.037
Test Batch ID	Male courtship	0.153 ± 0.052	---	0.613 ± 0.170	---	---
	Male EPE		---	---	---	---
	Female UP response			0.082 ± 0.030	---	---
	Female EP response				---	---
	Female EPP					---
Permanent environment	Male courtship	0.223 ± 0.028	0.450 ± 0.356	---	---	---
	Male EPE		0.080 ± 0.060	---	---	---
	Female UP response			0.201 ± 0.031	0.384 ± 0.266	0.284 ± 0.321
	Female EP response				0.034 ± 0.014	0.502 ± 0.198
	Female EPP					0.084 ± 0.033
Year	Male courtship	---	---	---	---	---
	Male EPE		0.099 ± 0.143	---	-0.035 ± 0.583	---
	Female UP response			---	---	---
	Female EP response				0.184 ± 0.178	---
	Female EPP					---
Clutch ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				---	---
	Female EPP					0.371 ± 0.030
Pair ID	Male courtship	---	---	---	---	---
	Male EPE		---	---	---	---
	Female UP response			---	---	---
	Female EP response				0.175 ± 0.039	---
	Female EPP					0.173 ± 0.031
Residual	Male courtship	0.420 ± 0.028	---	---	---	---
	Male EPE		0.772 ± 0.137	---	---	---
	Female UP response			0.508 ± 0.021	---	---
	Female EP response				0.562 ± 0.123	---
	Female EPP					0.271 ± 0.015

Table S14. Animal model IX, intrasexual, initial + new data, VCE, raw data. Like Table S6, but examining within-sex correlations among female traits only. Note that 'Pair ID' refers to the social pair for the traits 'WP response' and 'EPP' and refers to the combination of female and extra-pair male for the trait 'EP response'. UP = unpaired, EP = extra-pair, WP = within-pair.

		Fecundity	UP response	EP response	WP response	EPP
Genetic	Fecundity	0.192 ± 0.044	0.189 ± 0.194	0.590 ± 0.131	0.351 ± 0.151	0.031 ± 0.219
	UP response		0.221 ± 0.031	0.588 ± 0.143	0.050 ± 0.257	0.482 ± 0.335
	EP response			0.079 ± 0.021	0.298 ± 0.239	0.811 ± 0.158
	WP response				0.024 ± 0.009	0.014 ± 0.319
	EPP					0.034 ± 0.020
Test Batch ID	Fecundity	---	---	---	---	---
	UP response		0.119 ± 0.010	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					---
Experiment	Fecundity	0.114 ± 0.013	---	---	---	---
	UP response		---	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					---
Permanent environment	Fecundity	0.257 ± 0.047	-0.015 ± 0.134	-0.465 ± 0.190	0.077 ± 0.099	-0.322 ± 0.093
	UP response		0.172 ± 0.025	0.134 ± 0.188	0.136 ±0.146	0.351 ± 0.202
	EP response			0.055 ± 0.016	0.053 ± 0.143	0.780 ± 0.183
	WP response				0.078 ± 0.010	-0.215 ± 0.143
	EPP					0.109 ± 0.026
Year	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			0.028 ± 0.003	0.416 ± 0.041	---
	WP response				0.054 ± 0.005	---
	EPP					---
Clutch ID	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					0.209 ± 0.023
Pair ID	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			0.111 ± 0.006	---	---
	WP response				0.112 ± 0.004	---
	EPP					0.364 ± 0.014
Residual	Fecundity	0.437 ± 0.020	---	---	---	---
	UP response		0.488 ± 0.015	---	---	---
	EP response			0.727 ± 0.017	---	---
	WP response				0.732 ± 0.007	---
	EPP					0.283 ± 0.009

Table S15. Animal model X, intrasexual, initial + new data, MCMCglmm, raw data. Like Table S14, but model performed using MCMCglmm.

		Fecundity	UP response	EP response	WP response	EPP
Genetic	Fecundity	0.108 ± 0.097	0.296 ± 0.339	0.235 ± 0.255	0.160 ± 0.233	0.054 ± 0.310
	UP response		0.220 ± 0.039	0.410 ± 0.163	0.126 ± 0.213	0.241 ± 0.257
	EP response			0.045 ± 0.011	0.178 ± 0.179	0.504 ± 0.149
	WP response				0.026 ± 0.007	-0.022 ± 0.204
	EPP					0.087 ± 0.027
Test Batch ID	Fecundity	---	---	---	---	---
	UP response		0.124 ± 0.048	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					---
Experiment	Fecundity	0.093 ± 0.053	---	---	---	---
	UP response		---	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					---
Permanent environment	Fecundity	0.334 ± 0.107	-0.100 ± 0.233	-0.145 ± 0.204	0.160 ± 0.202	-0.333 ± 0.239
	UP response		0.169 ± 0.029	0.242 ± 0.208	0.094 ±0.227	0.283 ± 0.215
	EP response			0.034 ± 0.009	0.049 ± 0.193	0.508 ± 0.137
	WP response				0.025 ± 0.007	-0.104 ± 0.218
	EPP					0.118 ± 0.032
Year	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			0.061 ± 0.036	0.180 ± 0.377	---
	WP response				0.088 ± 0.065	---
	EPP					---
Clutch ID	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					0.350 ± 0.021
Pair ID	Fecundity	---	---	---	---	---
	UP response		---	---	---	---
	EP response			0.202 ± 0.013	---	---
	WP response				0.123 ± 0.015	---
	EPP					0.171 ± 0.025
Residual	Fecundity	0.465 ± 0.046	---	---	---	---
	UP response		0.486 ± 0.029	---	---	---
	EP response			0.658 ± 0.027	---	---
	WP response				0.737 ± 0.052	---
	EPP					0.275 ± 0.012

Table S16. Animal model XI, intrasexual, initial + new data, VCE, BLUPs. Like Table S14, but based on estimates of individual average phenotypes (using BLUPs). Here, the decomposition of variance is only into an additive genetic component, maternal effect component, and a residual component. Note that the heritability estimates (highlighted in grey) refer to the heritability of average phenotypes rather than the heritability of single measures (as is the case for models using raw data).

Effects		Fecundity	UP response	EP response	WP response	EPP
Genetic	Fecundity	0.100 ± 0.053	0.800 ± 0.240	0.612 ± 0.127	0.480 ± 0.272	0.304 ± 0.160
	UP response		0.347 ± 0.048	0.564 ± 0.180	0.003 ± 0.354	0.426 ± 0.209
	EP response			0.285 ± 0.041	0.532 ± 0.276	0.915 ± 0.079
	WP response				0.119 ± 0.044	0.231 ± 0.256
	EPP					0.128 ± 0.032
Maternal	Fecundity	0.067 ± 0.041	-1.000 ± 0.005	---	---	---
	UP response		0.058 ± 0.023	---	---	---
	EP response			---	---	---
	WP response				---	---
	EPP					---
Residual	Fecundity	0.835 ± 0.065	-0.049 ± 0.063	-0.103 ± 0.054	0.056 ± 0.073	0.226 ± 0.067
	UP response		0.595 ± 0.049	0.040 ± 0.076	0.073 ± 0.077	0.078 ± 0.065
	EP response			0.715 ± 0.041	0.006 ± 0.040	0.236 ± 0.045
	WP response				0.881 ± 0.044	-0.157 ± 0.057
	EPP					0.872 ± 0.032

147 **Table S17. Animal model XII, intrasexual, initial + new data, MCMCglmm, BLUPs.** Like Table
148 S16, but model performed using MCMCglmm.
149

Effects		Fecundity	UP response	EP response	WP response	EPP
Genetic	Fecundity	0.161 ± 0.082	0.176 ± 0.199	0.204 ± 0.163	0.018 ± 0.088	0.572 ± 0.311
	UP response		0.399 ± 0.055	0.334 ± 0.107	0.031 ± 0.089	0.431 ± 0.229
	EP response			0.565 ± 0.062	0.038 ± 0.077	0.393 ± 0.165
	WP response				0.541 ± 0.030	0.012 ± 0.088
	EPP					0.119 ± 0.111
Residual	Fecundity	0.839 ± 0.082	---	---	---	---
	UP response		0.601 ± 0.055	---	---	---
	EP response			0.435 ± 0.062	---	---
	WP response				0.459 ± 0.030	---
	EPP					0.881 ± 0.111

150
151

Table S18. Selection line descriptive statistics. Basic statistics for each of the 6 selection lines (columns) across four generations ('S0' = generation of founders, 'S1' to 'S3' subsequent offspring generations). Indicated are the numbers of birds, their mean inbreeding coefficient F, their average courtship rate and standard deviation SD (units are square-root transformed number of seconds), the number of males measured, the number of courtship trials, the average age of males in days during the courtship trials, and numbers of males and females that participated in the breeding experiments (for measurement of extra-pair paternity levels). The rightmost column shows either the total sum (for sample sizes) or the grand mean (average of six lines).

	low 1	low 2	control 1	control 2	high 1	high 2	total
S0 birds to choose from	693	693	773	773	693	693	773
S0 parents producing S1	38	39	35	36	36	39	223
S1 offspring produced	114	86	116	88	79	85	568
S1 birds to choose from	109	81	110	94	72	80	546
S1 parents producing S2	31	32	30	34	29	29	185
S2 offspring produced	84	87	102	141	105	111	630
S2 birds to choose from	80	85	92	130	94	103	584
S2 parents producing S3	26	29	27	27	28	26	163
S3 offspring produced	76	141	127	118	127	92	681
S1 average realized F	0.002	0.021	0.025	0.019	0.023	0.040	0.022
S2 average realized F	0.065	0.060	0.054	0.077	0.064	0.080	0.067
S3 average realized F	0.105	0.110	0.116	0.121	0.125	0.124	0.117
S1 mean courtship rate	2.651	2.414	3.006	2.626	3.897	3.998	3.098
S2 mean courtship rate	1.635	2.027	3.363	2.623	4.158	3.804	2.935
S3 mean courtship rate	1.416	1.619	3.578	3.381	4.695	4.678	3.228
S1 between-individual SD in courtship rate	1.678	1.584	1.704	1.401	1.523	1.906	1.633
S2 between-individual SD in courtship rate	1.456	1.528	1.607	1.635	1.635	1.825	1.614
S3 between-individual SD in courtship rate	1.336	1.285	1.263	1.511	1.397	1.342	1.356
S1 males measured for courtship rate	65	44	58	45	42	37	291
S2 males measured for courtship rate	45	47	56	61	47	47	303
S3 males measured for courtship rate	40	63	62	51	53	34	303
S1 courtship rate trials	260	174	230	180	168	144	1156
S2 courtship rate trials	178	188	220	242	188	188	1204
S3 courtship rate trials	156	252	248	204	212	136	1208
S1 mean male age during trials [a]	164.9	161.4	163.2	162.1	163.9	161.1	162.8
S2 mean male age during trials [a]	165.5	164.8	165.9	165.8	167.4	164.9	165.7
S3 mean male age during trials [a]	167.3	169.5	167.5	167.3	168.5	165.7	167.6
S3 males tested in breeding experiment	30	42	36	36	41	32	217
S3 females tested in breeding experiment	25	47	37	37	37	36	219

[a] Males had two courtship trials at the age of about 120 days (range for 'S1-S3': 100-140) and another two trials at the age of about 210 days (range for 'S1-S3': 188-228).

General discussion

My PhD thesis focused on mate choice in monogamous species. Proximately, I studied male mate choice (**Chapter 1**) and mutual mate choice (**Chapter 2**) using classic methods of behavioral ecology; meta-analytically, I studied male attractiveness (**Chapter 3**) and assortative mating in birds (**Chapter 4**); evolutionarily, I studied female promiscuity (female extra-pair mate choice) using an explicit quantitative genetics approach (**Chapter 5**). In the following, I will highlight important results and connect these five chapters throughout my thesis. Furthermore, I envisage some directions for future research.

Mate choice from different sexual perspectives

Male mate choice for female fecundity

Systematically studying lifetime monogamous species from different sexual perspectives could gain an overall picture of mate choice (Edward and Chapman 2011; Courtiol et al. 2016). Given a massive literature focusing on female mate choice, male mate choice necessitates more attention (Edward and Chapman 2011). In **Chapter 1**, we carried out an experiment in which we measured the variation of female fecundity, and tested whether males preferred unfamiliar females with high fecundity (30 eggs laid on average) over those of low fecundity (6 eggs laid on average). Zebra finch males preferred the high-fecundity female in 59% of choice tests that lasted 20 min. When extending such choice tests over several days, male "success" in associating with the high-fecundity female was still modest (61% correct choices). Overall, male zebra finches seem to have only limited abilities to identify the better mate when faced with a choice between extremes in terms of female fecundity. We speculate that such a preference may not have evolved because predicting fecundity from female phenotypes (e.g. body weight) is not sufficiently accurate given the small amount of fitness variation explained by these phenotypic traits. Future studies investigating reliability of the assumed quality indicators (to which degree the indicator could explain the individual fitness, see **Chapter 2**) seem meaningful. Costs of choosing (such as intra-sexual competition, out of mating pool after paired, and time and sampling effort to identify a high-quality partner, **Chapter 2**; Dechaume-Moncharmont et al. 2016) relative to benefits of choosing also need be accounted for explaining the evolution of choosiness.

Assortative mating for fitness (mutual mate choice) in zebra finches

In species with bi-parental care, mutual mate choice is expected to result in assortative mating for quality (Jones and Hunter 1993). By combining multiple measures of causes (inbreeding, early nutrition) and consequences (ornaments, displays, fitness components) of variation in quality into a single principal component, in **Chapter 2**, we showed that quality variation can be quantified successfully. We further showed that variation in quality indeed predicts individual pairing success, presumably because it reflects an individual's vigour or ability to invest in reproduction. However, despite high statistical power, we found no evidence for assortative mating.

To expect a positive correlation arises from mutual mate choice, there would have to be a reasonably high degree of between-individual agreement about what constitutes an attractive partner. For instance, in humans, body height is one of the criteria for mate choice. Directional preferences and mutual mate choice of this trait result in a Pearson correlation coefficient of r = 0.23 (meta-analysis of 154 estimates; Stulp et al. 2017). Yet, in zebra finches, both the male and the female show only low levels of such agreement (**Chapter 1**; Forstmeier and Birkhead 2004), meaning that their preferences are highly individualistic or flexible. Such individual mating preferences imply that intra-sexual competition for the presumed highest-quality mates is reduced and that most individuals may be able to pair with their preferred mate and achieve maximal fitness through effective cooperation in a lifelong pair bond (Ihle et al. 2015). In contrast, under the conventional scenario with consensus in mate preferences, all members of a sex compete for the same (few) high-quality partners, i.e. for those potential mates that show the highest values for quality indicator traits. Under such conditions, most if not all individuals would pay a cost for the intense competition, while the successful competitors only achieve a relatively small benefit from being choosy, unless the quality indicators strongly and reliably predict fitness. Moreover, unsuccessful competitors would end up unpaired or paired to non-preferred mates, which may result in unstable partnerships that suffer (in terms of fitness) from a lack of mutual commitment. Thus, at a high level such as group selection in monogamous species, individualistic mate choice seems more stable. Theoretical modelling and empirical evidence are required in future studies in this respect. Specifically, studies that highlight (1) the degree of honesty of signals (e.g. how well expression of the signal reflects individual quality), (2) the strength of preferences for these signals, and (3) the costs of competition for mates are promising.

Assortative mating in birds: is mutual mate choice common?

According to literature, assortative mating is not only common in humans but also ubiquitous in the animal kingdom (Jiang et al. 2013). To further reveal whether the observed assortment results primarily from the evolution of mutual mate choice ('likes-attract

hypothesis', Buss 1985) or from confounding ecological factors and estimation bias, in **Chapter 4**, we carried out a meta-analysis by extracting effect sizes of assortative mating from published literature. In order to disentangle the publication bias, we did comparisons of effect sizes between published and long-term unpublished data. The conventional literature search yielded published estimates of assortative mating (r = 0.201) that were higher than unpublished estimates from our own long-term field studies (r = 0.106), reflecting the inevitable publication- and ascertainment bias. Second, the unpublished correlations were significantly affected by shared observer-error as well as by temporal and spatial autocorrelation. Third, we found no assortative mating for size in the only species that has been studied experimentally where all sources of bias are excluded (r = -0.003). Hence, the ubiquity of assortative mating probably results from multiple confounding factors and not because 'likes attract' (**Chapter 2**). Those confounding factors such as temporal and spatial autocorrelation deserved closer investigation because of their own ecological meanings (Valcu and Kempenaers 2010; Rolan-Alvarez et al. 2015). On the other hand, to avoid confounding factors such as publication bias, we need to shift our efforts towards unbiased quantification of some key parameters (Parker et al. 2016). This would require complete and unbiased reporting of all relevant parameters that have been examined, or at least reporting an unbiased estimate of the average effect size within a study (Forstmeier et al. 2016).

Male attractiveness, female promiscuity

The effects of color bands on male attractiveness

For decades, researchers working with birds have individually marked their study species with colored leg bands. In behavioral ecology, the hypothesis that colorful leg bands can alter the attractiveness of male or female zebra finches (Burley et al. 1982), with resulting effects on behavior (Burley 1986), physiology (Gil et al. 1999), life-history and fitness (Burley 1985; Burley et al. 1996), have been quite influential. However, our observation that zebra finch mate choice seems predominantly individual specific rather than following a universal rule of attractiveness (Forstmeier and Birkhead 2004; Ihle et al. 2015; **Chapter 1; Chapter 2**) is at odds with the existence of universal band-color effects on attractiveness. Moreover, many fields of science – including behavioral ecology – are currently experiencing a heated debate about the extent to which publication bias against null-findings results in a misrepresentative scientific literature. Enlightened by points mentioned above, in **Chapter 3**, under a framework of meta-analysis, we carried out a conceptional replication using data with multiple populations from multiple labs. 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. This is 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. Again, our results were consistent with previous studies (i.e. individualistic mate choice, **Chapter 1; Chapter 2**). We argue that our field – and science in general – would benefit from more effective means to counter confirmation bias and publication bias.

Genetic constraints of female promiscuity

Finding out the genetic architecture of complicated behavior such as female extra-pair mating is appealing. In **Chapter 5**, according to theoretical hypothesis and previous empirical evidence, we used an explicit quantitative genetic approach with well-designed experiments (e.g. selection lines, mating with different partners) to assess two evolutionary forces of female extra-pair mating behavior. We found, evolutionarily, these two evolutionary forces (inter-sexual and intra-sexual) both exist but the drive of intra-sex is relative larger. Proximately, we found the social environment around the focal female (her social partner, the potential extra-pair mating males) could affect her extra-pair mating behavior substantially.

Summary

The key results of my thesis revealed that zebra finches tend to choose social mating partner individualistically. This strategy of mate choice is evolutionarily stable because every individual in that mating pool could find a preferred mate. This will result in maximized fitness of the population. In contrast, the pattern of mate choice in lekking species in which only a few high quality individuals could mate and contribute genes to next generation (small size of effect population) seems to be an evolutionarily unstable strategy to monogamous species. Our results are also consistent with previous findings in this species (see Forstmeier and Birkhead 2004; Ihle et al. 2015) which highlighted individualistic mate choice and the behavioral compatibility between pair partners. Future studies focusing on traits that function as individual marking such as highly individual song are very promising giving the individualistic mate choice in this species.

In zebra finches, males show four additional plumage ornaments: orange cheek patches, a black breast band, fine black and white stripes on the chin and upper breast, and chestnut-brown flanks with white spots (Jeronimo et al. 2018). Given the apparent sexual dimorphism of this species (Figure 1), people still wonder the mechanism underlying this sexual dimorphism and expect that the ornaments of male zebra finches are under sexual selection. We have presented some evidence **(Chapter 2)** to show these traits may be not sexually selected traits. Furthermore, phylogenetically, these ornaments are ancestral traits because of their occurrence in other relative species as well. For example, the white spots on chestnut-brown flanks exists in species such as the double-barred finch, *Taeniopygia*

bichenovii, painted finch, *Emblema pictum*, red-eared firetail, *Stagonopleura oculata* and diamond firetail *Stagonopleura guttata*. The black breast appears in long-tailed finch *Poephilia acuticauda* and black-throated finch *Poephilia cincta* as well (Jetz et al. 2012; Singhal et al. 2015). However, all these four male ornaments together is an unique combination. Therefore, zebra finches can be recognized easily from other species. Finally, all these male ornaments explained limited variation of male fitness and were not important to female choice (Jeronimo et al. 2018). Therefore, currently, these ornaments of male zebra finches may function as a way of species recognition but not signals of sexual selection.

Conclusion

Despite great efforts contributed to research of sexual selection since Darwin, the key aspects of sexual selection and mate choice are still plagued by confusion and disagreement. Many of these areas are complex and require new theory and empirical data for complete resolution (Jones and Ratterman 2009). My thesis studied mate choice in zebra finches and found that this species has a pattern of individualistic mate choice but not choice with consensus of a few high-quality individuals (**Chapter 1, 2, 3**). In this respect, a more details-monitoring way (e.g. initial mate preference, consequently mate choice, copulation attempts, pairing bond forming, coordination after pairing, and divorce) seems promising.

Another main finding of my thesis is the irreproducibility of key findings from previous studies. This situation is not only happening in behavioral ecology only (Ioannidis 2005; Begley and Ellis 2012; Button et al. 2013; Open-Science-Collaboration 2015; Baker 2016; Forstmeier et al. 2016; Parker et al. 2016). Thus, the value of replication in terms of key findings in one research field seems appealing and apparent. Perhaps, more specifically, as a PhD student or young scientist, the main point to keep in mind is that the limitations of a study shouldn't be hidden, but openly acknowledged. Taking the example of my thesis, the limitations of the zebra finch model for understanding mate choice might lie in the difference between captive and natural environment.

There are many reasons behind the irreproducibility of studies. For instance, the limited statistical power because of small sample sizes; multiple tests with selective reporting; hypothesizing after the results are known ('HARKing'); pseudoreplication at different levels; publication bias plus other cognitive biases such as optimistic bias of the experimenters, observation without blinding and false positive finding due to data structure or neglecting other important factors (Kerr 1998; Forstmeier et al. 2016; Parker et al. 2016; **Chapter 3, 4, 5**). In order to avoid those issues listed above, solutions such as preregistration of studies, replication and rigorous assessment of context dependence for a more general pattern and

blinding during the data collection are recommended (Forstmeier et al. 2016; Parker et al. 2016; Ihle et al. 2017).

References

Baker, M. 2016. Is There a Reproducibility Crisis? Nature 533:452-454.

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

Burley, N. 1985. Leg-Band Color and Mortality Patterns in Captive Breeding Populations of Zebra Finches. Auk 102:647-651.

Burley, N. 1986. Sexual selection for aesthetic traits in species with biparental care. The American Naturalist 127:415-445.

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

Buss, D. M. 1985. Human mate selection: Opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable. Am Sci 73:47-51.

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 (vol 14, pg 365-376, 2013). Nat Rev Neurosci 14:444-444.

Courtiol, A., L. Etienne, R. Feron, B. Godelle, and F. Rousset. 2016. The Evolution of Mutual Mate Choice under Direct Benefits. Am Nat 188:521-538.

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.

Edward, D. A. and T. Chapman. 2011. The evolution and significance of male mate choice. Trends Ecol Evol 26:647-654.

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., E. J. Wagenmakers, and T. H. Parker. 2016. Detecting and avoiding likely false‐positive findings–a practical guide. Biological Reviews.

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

Ihle, M., B. Kempenaers, and W. Forstmeier. 2015. Fitness Benefits of Mate Choice for Compatibility in a Socially Monogamous Species. Plos Biol 13.

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.

Ioannidis, J. P. A. 2005. Why most published research findings are false. Plos Med 2:696-701.

Jeronimo, S., M. Khadraoui, D. P. 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 29:459-467.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444-448.

Jiang, Y. X., D. I. Bolnick, and M. Kirkpatrick. 2013. Assortative Mating in Animals. Am Nat 181:E125-E138.

Jones, A. G. and N. L. Ratterman. 2009. Mate choice and sexual selection: What have we learned since Darwin? P Natl Acad Sci USA 106:10001-10008.

Jones, I. L. and F. M. Hunter. 1993. Mutual Sexual Selection in a Monogamous Seabird. Nature 362:238-239.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review 2:196-217.

Open-Science-Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349:aac4716.

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.

Rolan-Alvarez, E., A. Carvajal-Rodriguez, A. de Coo, B. Cortes, D. Estevez, M. Ferreira, R. Gonzalez, and A. D. Briscoe. 2015. The scale-of-choice effect and how estimates of assortative mating in the wild can be biased due to heterogeneous samples. Evolution 69:1845-1857.

Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Strand, Q. Y. Li, B. Raney, C. N. Balakrishnan, S. C. Griffith, G. McVean, and M. Przeworski. 2015. Stable recombination hotspots in birds. Science 350:928-932.

Stulp, G., M. J. P. Simons, S. Grasman, and T. V. Pollet. 2017. Assortative mating for human height: A meta-analysis. Am J Hum Biol 29.

Valcu, M. and B. Kempenaers. 2010. Spatial autocorrelation: an overlooked concept in behavioral ecology. Behav Ecol 21:902-905.

Wang, D., W. Forstmeier, M. Ihle, M. Khadraoui, S. Jerónimo, K. Martin, and B. Kempenaers. 2018. Irreproducible text‐book “knowledge”: The effects of color bands on zebra finch fitness. Evolution.

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

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

Addresses of co-authors

Tomáš Albrecht
Institute of Vertebrate Biology of the Czech Academy of Sciences, 603 65, Brno, and Faculty of Science, Charles University, 128 44 Prague, Czech Republic
Email: albrecht@ivb.cz

Christiaan Both
Conservation Ecology Group, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands
Email: c.both@rug.nl

Martin Bulla
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany

NIOZ Royal Netherlands Institute for Sea Research, Department of Coastal Systems and Utrecht University, P.O. Box 59, 1790 AB Den Burg, The Netherlands

Department of Ecology, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21 Prague 6 - Suchdol, Czech Republic
Email: bulla.mar@gmail.com

Niels Dingemanse
Behavioural Ecology, Department of Biology, Ludwig Maximilians University of Munich, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany
Email: dingemanse@biologie.uni-muenchen.de

Renée A. Duckworth
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA
Email: rad3@email.arizona.edu

Wolfgang Forstmeier
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: forstmeier@orn.mpg.de

Malika Ihle
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany

Department of Entomology and Nematology, University of Florida, 1881 Natural Area Dr.,Gainesville, FL 32611, USA
Email: malika_ihle@hotmail.fr

Sofia Jerónimo
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: sofiajdossantos@gmail.com

Patrik Karell
Bioeconomy Research Team, Novia University of Applied Sciences, Raseborgsvägen 9, FI-10600 Ekenäs, Finland
Email: Patrik.Karell@novia.fi

Bart Kempenaers
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: b.kempenaers@orn.mpg.de

Nele Kempenaers
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany

Email: nelekempenaers@outlook.com

Mehdi Khadraoui
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: mehdi.khadraoui@outlook.com

Lynna Marie Kiere
Department of Evolutionary Ecology, Universidad Nacional Autónoma de México, Mexico City, Distrito Federal, Mexico

Center for the Study of Biodiversity and Conservation (CIByC), Autonomous University of the State of Morelos, Cuernavaca Morelos, Mexico
Email: lynnamiss@gmail.com

Katrin Martin
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: k.martin@orn.mpg.de

Mihai Valcu
Department of Behavioural Ecology & Evolutionary Genetics
Max Planck Institute for Ornithology
Seewiesen, Germany
Email: valcu@orn.mpg.de

Alastair Wilson

Centre of Ecology and Conservation, College of Life and Environmental Sciences,

University of Exeter, Penryn Campus,

Penryn, Conrwall, TR10 9FE, UK

Email: A.Wilson@exeter.ac.uk

Author contributions

Chapter 1: W.F. conceived the study. D.W. and N.K. collected data. W.F. and D.W. analyzed the data with input from B.K. D.W. and W.F. wrote the first draft of the manuscript. All authors read and commented on the first draft of the manuscript. All authors approved the final manuscript.

Chapter 2: W.F., B.K. and D.W. contributed to the design of the research. D.W. and W.F. measured ornaments and recorded songs. D.W. contributed to nest checks and observations of pairing and parental care. W.F. and D.W. analyzed the data. D.W., W.F. and B.K. wrote the paper.

Chapter 3: W.F., D.W., and B.K. conceived the project. D.W. and W.F. analyzed the data. B.K., D.W., and W.F. wrote the manuscript with contribution of M.I. and M.K. Data collection was carried out by M.K., S.J. (experiments 1, 2, and 5), M.I. (experiments 3 and 4), K.M. (experiment 6), and D.W. (experiment 8).

Chapter 4: D.W. and W.F. conceived the project with input of B.K. W.F., M.V. and D.W. analyzed the data. D.W. and W.F. coordinated the project with input from N.D.. T.A. contributed the data of Barn Swallow. L. M. K. contributed the data of Blue-footed Booby. B.K. and M.V. contributed the data of Blue Tit. N.D. contributed the data of Great Tit. C.B. contributed the data of Pied Flycatcher. B.K. and M.B. contributed the data of Semipalmated Sandpiper. P.K. contributed the data of Tawny Owl. R.A.D. contributed the data of Western Bluebirds. B.K., W.F. and D.W. contributed the data of Zebra Finch. D.W. and W.F wrote the first draft. All coauthors contributed their effort for writing the manuscript. D.W., W.F and M.B. integrated the comments from the other co-authors. B.K. wrote the final version with input from D.W., W.F. and M.B..

Chapter 5: W.F., D.W. and B.K. conceived the study. D.W., W.F. and K.M. collected the data. D.W. and W.F. analyzed the data. D.W. wrote the manuscript with input from W.F., B.K. and A.W..

…….
Bart Kempenaers
Doktorvater

…….
Daiping Wang
Doktorand

Acknowledgments

My deepest thanks go to Wolfgang Forstmeier. He changed me from a naïvely new recruit to an advanced PhD student. First, he is intelligent with the scientific breadth and unconventional-thinking which inspired me deeply. Further, as a supervisor, he is extremely patient and always comforted me when I was anxious about my projects. He used an excellent way to teach me how to do good-quality rather than quick and dirty science in times of an ever increasing pressure to publish significant results. In our zebra finch group, it is a consensus that he is the best supervisor we can imagine. Beside academic issues, Wolfgang is the 'Doktorvater' of my daily life in Germany as well. As a Chinese student who never studied abroad before, he helped me a lot get used to living in Germany. I really enjoyed very much the times we had dinner together. Our daily discussions (scientific or unscientific) were always exciting and intriguing. I will miss them.

Second, I want to thank Bart Kempenaers for his excellent supports to my research. With his broad vision of my research field, he provided outstanding supervision to me. He supported me in all my scientific projects and enabled me to carry out my own ideas. He is an excellent scientific writer which is especially helpful. Thank you very much for your faith in me.

My thanks go to my PhD advisory committee member Alastair Wilson who gave me encouragement and professional advice of quantitative genetics. I also want to thank my co-authors: Nele Kempenaers for practical work; Mihai Valcu helped me in issues of statistics; Martin Bulla helped me to improve the level of manuscripts; Niels Dingemanse, Christiaan Both, Renée A. Duckworth, Lynna Marie Kiere, Patrik Karell, Tomáš Albrecht, Malika Ihle, Mehdi Khadraoui and Sofia Jerónimo for their contribution of data. Finally, I thank Carol Gilsenan for proofreading of my thesis.

My thanks with all my hearts go to the two exceptional and tireless technical assistants: Katrin Martin and Melanie Schneider. They have done the practically tedious work with the birds and in the lab. Katrin Martin, I am so grateful for your efforts: breeding the birds, the daily nest checks, the daily behavioral observations, maintenance of the databases and scoring of the thousands of videos. Melanie Schneider, I thank you for your efficient and reliable lab work resulting in parentage analysis of thousands of birds.

My work would not have been possible without a large group of care takers looking after the birds every day: Sonja Bauer, Edith Bodendörfer, Jane Didsbury, Annemarie Grötsch, Andrea

Kortner, Frank Lehmann, Petra Neubauer, Katharina Piehler, Frances Weigel and Barbara Wörle. I am very grateful for your work.

The wonderful and friendly working environment of the department Kempenaers is a premise for a productive PhD. For this, I would like to thank (those who not mentioned above) Christine Baumgartner, Kristina Beck, Luisana Carballo, Carmen Dobus, Pamela Espíndola Hernández, Carol Gilsenan, Helga Gwinner, Elena Kappers, Ulrich Knief, Johannes Krietsch, Sylvia Kuhn, Peter Loës, Christina Muck, Jakob C. Mueller, Yifan Pei, Peter Santema, Emmi Schlicht, Lotte Schlicht, Peter Skripsky, Kim Teltscher, Agnes Türk, Cristina-Maria Valcu and Andrea Wittenzellner.

The daily life in Starnberg (Ghetto) was and still is important to me and my PhD projects. For this, I would like to thank Malika Ihle, Ulrich Knief, Yifan Pei, Shouwen Ma and his family, Peter Santema, Johannes Krietsch, Kristina Beck, Carol Gilsenan, Elena Kappers, Peter Skripsky, Luke Eberhart-Phillips, Meng-Ching Ko, Qiaoyi Liang, Ignas Safari Mng'anya and Amanda de Almeida Monte.

The football team of 'Seewiesen United' which is always a lot of fun, the team members are: Wolfgang Forstmeier, Mauricio Nicolas Adreani, Luke Eberhart-Phillips, Peter Santema, Johannes Krietsch, Elena Kappers, Pietro D'Amelio, Lucía Mentesana, Klaus Pichler and Felix Hartl.

Last, I am very grateful for my parents and other family members who supported me as always.

Daiping Wang
Seewiesen
August 2018

Curriculum vitae

Daiping Wang

Max Planck Institute for Ornithology,
Eberhard-Gwinner-Strasse, 82319 Seewiesen
Tel: +49-8157-932262 Email: dwang@orn.mpg.de
Date of Birth: 15st September, 1988

Education

- **2014 – to date: PhD student, International Max Planck Research School for Organismal Biology (IMPRS), Max Planck Institute for Ornithology, Seewiesen, Germany**
  Thesis: Mate choice and the evolution of female promiscuity in a social monogamous species
- **2011 – 2014: Master in animal ecology (zoology, ornithology, ecology), Beijing normal University, Beijing, China**
  master-thesis: The activity and function of dawn chorus in brownish-flanked bush warber (*Cettia fortipes*)
- **2007 – 2011: Bachelor in biology, Hainan Normal University, Haikou, China**

Publications

Peer-reviewed

1. Jeronimo, S., M. Khadraoui, **D. P. 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* 29:459-467.

2. **Wang, D.**, W. Forstmeier, M. Ihle, M. Khadraoui, S. Jerónimo, K. Martin, and B. Kempenaers. 2018. Irreproducible text‐book 'knowledge' : The effects of color bands on zebra finch fitness. *Evolution*. 72(4):961-976.

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

4. **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.

5. Knief, U., W. Forstmeier, Y. F. Pei, M. Ihle, **D. P. Wang**, K. Martin, P. Opatova, J. Albrechtova, M. Wittig, A. Franke, T. Albrecht, and B. Kempenaers. 2017. A sex-chromosome inversion causes strong overdominance for sperm traits that affect siring success. *Nature Ecology & Evolution* 1:1177-1184.

6. **Wang, D.**, C Xia, Y Liu, Y Sun. 2017. The theory of quantitative genetics into evolutionary ecological studies: the applications of Animal Model. *Chinese Journal of Zoology*, 52(4): 690-701.

7. Wei, C. T., C. X. Jia, L. Dong, **D. P. Wang**, C. W. Xia, Y. Y. Zhang, and W. Liang. 2015. Geographic variation in the calls of the Common Cuckoo (*Cuculus canorus*): isolation by distance and divergence among subspecies. J Ornithol 156:533-542.

8. Xia, C. W., B. Y. Liu, **D. P. Wang**, H. Lloyd, and Y. Y. Zhang. 2015. Reliability of the Brownish-flanked Bush Warbler's soft song in male-male conflict. *Avian Res* 6.

9. **Wang, D.**; Xia, C.; Zhang, L.; Liu, J.; Liu, B.; Zhang, Y.: A simulation experiment on acoustic location system of wildlife and its accuracy analysis. *Chinese Journal of Zoology*, 2013, 5, 726-731.

10. **Wang, D.**; Huang, X.; Xia, C.: Seicercus soror: a bird new in Beijing province, China. Chinese *Journal of Zoology*, 2010, 219.

11. Rao, X.; D. X. Chen, **D. P. Wang**, Y. D. Li, W. Liang: A review on the study of relationship between birds and seed dispersal. *Journal of Anhui Agri. Sci.*, 2010, 38(26): 14209-14212.

Talks

- European society for evolutionary biology (ESEB), Groningen, Netherlands (2017)
- 13th China Ornithological Congress, Hefei, China (2015)
- 11th Chinese graduate students kingfisher forum, Hefei, China (2015)
- 9th Chinese graduate students kingfisher forum, Hangzhou, China (2013)
- IMPRS retreat (2016)

Posters

- The 16th International Behavioral Ecology Congress (ISBE), Exeter, England (2016)

Awards

- 2017: IMPRS for organismal biology “best student paper award” (1st)

Statutory declaration and statement

Eidesstattliche Versicherung

Ich versichere hiermit an Eides statt, dass die von mir vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist.

München, den 7.8.2018 Daiping Wang

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist. Im Weiteren erkläre ich, dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe oder ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

München, den 7.8.2018 Daiping Wang