VOLUME 81, NO. 3 SEPTEMBER 2006

THE QUARTERLY REVIEW *of* BIOLOGY

REPLICATING EMPIRICAL RESEARCH IN BEHAVIORAL ECOLOGY: HOW AND WHY IT SHOULD BE DONE BUT RARELY EVER IS

CLINT D. KELLY

Department of Biology, University of Toronto at Mississauga
Mississauga, Ontario L5L 1C6 Canada[1]

E-MAIL: CLINT.KELLY@ANU.EDU.AU

KEYWORDS

study replication, quasireplication, *p*-value, null hypothesis statistical testing, meta-analysis

If the data collected in the past say nothing about data to be gathered in the future, empirical research is merely historical. (Krueger 2001:21)

ABSTRACT

That empirical evidence is replicable is the foundation of science. Ronald Fisher, a founding father of biostatistics, recommended that a null hypothesis be rejected more than once because "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon" (Fisher 1974:14). Despite this demand, animal behaviorists and behavioral ecologists seldom replicate studies. This practice is not part of our scientific culture, as it is in chemistry or physics, due to a number of factors, including a general disdain by journal editors and thesis committees for unoriginal work. I outline why and how we should replicate empirical studies, which studies should be given priority, and then elaborate on why we do not engage in this necessary endeavor. I also explain how to employ various statistics to test the replicability of a series of studies and illustrate these using published studies from the literature.

[1] Present address: School of Botany and Zoology, Australian National University, Canberra ACT, Australia, 0200.

The Quarterly Review of Biology, September 2006, Vol. 81, No. 3

0033-5770/2006/8103-0001$15.00

THAT EMPIRICAL EVIDENCE is replicable is the foundation of science and the path to cumulative knowledge (Fisher 1974; Nickerson 2000). Consequently, readers of published research demand confidence that a paper's findings constitute what Fisher called a "demonstrable" phenomenon. Fisher (1974) recommended that a null hypothesis be rejected more than once because "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon" (p 14). In other words, by replicating a study and again obtaining a statistically significant result of $\alpha = 0.05$, we are able to safeguard against Type I error. For example, a one-time finding with a statistically significant result at the 0.05 level could have occurred by chance; however, the likelihood is small that a result at this level would be obtained repeatedly if the effect was, in fact, due to chance (Nickerson 2000).

Despite replication being the foundation of science, some have claimed that behavioral ecologists have done a poor job of replicating empirical studies (Coyne 1998; Palmer 2000; Birkhead 2002). Borrowing from Bacon (1926), Palmer (2000:442) lamented that ignoring replicate studies simply "perpetuates a collective contract of error" and greatly retards our efforts to achieve a robust understanding of evolutionary phenomena. I endeavor to determine if that claim is valid and, if so, why this is the case and how we can remedy the situation. I begin by defining replication and describing the different types of replication possible. Next, I quantitatively assess whether behavioral ecologists replicate, and explain why we should repeat studies. I then discuss in general terms how we should replicate empirical studies, which studies should be given priority for replication, and elaborate on why we do not engage in this necessary activity.

With this paper, I wish to introduce behavioral ecologists to the literature on study replication generated in other scientific disciplines, most notably psychology. My aim is to initiate discussion regarding study replication in our field and motivate researchers to repeat empirical research. Much discussion in the literature has recently focused on the relationship between the intensely competitive nature of bioscience and an apparent increase in scientific misconduct (Pearson 2003; Fenning 2004; Martinson et al. 2005; Anonymous 2006). Is behavioral ecology any different than the other biosciences? Perhaps we are not as competitive as the medical sciences, and so do not have the same problems with misconduct—but until we collect adequate data on the issue (see Montgomerie and Birkhead 2005), let us assume we are no different. To this end, I argue that a generous helping of study replication may ensure our field's health.

What Is and What Is Not Replication

true replication versus quasireplication

Behavioral ecologists often repeat studies with different species or systems. This has recently been termed "quasireplication" by Palmer (2000; see also Møller and Jennions 2001), and is not considered true replication. True replication arises in two contexts. First, in an estimation context, replication means getting exactly, or almost exactly, the same *effect* (direction and size) in an experiment in which experimental conditions (i.e., model species, design, and analysis) are very similar to the original study (Nickerson 2000). In other words, two point or interval estimates replicate one another when they meet some criterion of proximity (e.g., overlap of confidence intervals) (Greenwald et al. 1996; Colegrave and Ruxton 2003). Second, in a hypothesis testing context, two *statistical tests* replicate one another when they support the same conclusion (nonrejection or rejection in the direction specified by the authors) with respect to the same null hypothesis (Greenwald et al. 1996; Nickerson 2000). A large literature recommends the estimation approach, however, null hypothesis testing remains most popular (Greenwald et al. 1996).

types of true replication

True replication involves studying the same species whereas quasireplication uses a different species to the original study. Three different types of true replication are typically identified: (a) exact, (b) partial, and (c) conceptual (Lykken 1968; Hendrick 1991). Exact replication involves duplication of the first investigator's sampling procedure, experimental conditions, measuring techniques, and methods of analysis. Partial replication involves some change in procedure while other aspects are duplicated as in the original study. Conceptual replication involves investigating the same relationships or constructs as the original study, but in a procedurally different manner. In conceptually replicated studies, the investigators would use nothing more than the empirical finding from the original publication and formulate their own methods of sampling, measurement, and data analysis.

Exact replications that produce the same results increase confidence in the original study but have little novelty value, whereas successful conceptual and partial replications increase confidence in the original study *and* provide novelty by way of original data (Lykken 1968; Hendrick 1991). In contrast, unsuccessful conceptual and partial replications add nothing new and may also decrease confidence in the original study (Lykken 1968; Hendrick 1991). Therefore, "little or nothing can be concluded, so that unsuccessful [conceptual and partial] replications . . . are simply failures, at least with respect to the initial experiment" (Hendrick 1991:47). The value of a successful exact replication is generally determined by the importance of the original study's results but will yield "only grudging acknowledgement of helpfulness to the science because of lack of novelty value in the results" (Hendrick 1991:48). Hence, successful replication does not guarantee publication (Nickerson 2000). Unsuccessful exact replications, on the other hand, create an interesting circumstance. Hendrick (1991) considers unsuccessful exact replications "an annoyance" because they tend to "upset the equilibrium within the communication system of science" (p 47). He argues that no clear basis remains for deciding between the original and replicate study, which leaves the results of both as undefined and indeterminate (Hendrick 1991). Of course, we as readers and authors can decide for ourselves the quality of a paper and choose whether or not to cite it. Typically, however, in these cases we see the classic approach of citing a number of supporting studies and then stating "but see . . . " for contradictory evidence. Workers also tend to discount or ignore the unsuccessful replicate study if the original outcome is considered a "textbook example." This is particularly likely if the original study showed a positive outcome, and the replicate indicated no difference (Hendrick 1991). It is also possible that there is a "file drawer effect." In other words, the published literature represents a biased sample of the research actually conducted. This phenomenon could arise if a researcher chooses not to submit for publication a replicated study with a result contradictory to a hypothesis they wish to confirm, or if the study yields a statistically nonsignificant result. Recent direct evidence suggests that the statistical significance of a study plays a limited role in whether a manuscript is accepted for publication (e.g., Koricheva 2003; Møller et al. 2005); however, indirect evidence suggests otherwise (Gontard-Danek and Møller 1999; Palmer 1999; Jennions and Møller 2002).

Even if replicate studies are performed, the problem of these studies not being cited by subsequent researchers sill exists. There are several factors contributing to a study's neglect, including a subsequent author's decision that the replicated study is not pertinent to their own study, or the paper simply being missed in a literature search. This last reason should not be a factor today given the almost universal access to electronic literature databases (e.g., ISI Web of Science). An example of researchers ignoring contradictory evidence involves female mate preference for symmetrical males in zebra finches (*Taeniopygia guttata*). In a highly influential paper, Swaddle and Cuthill (1994) showed that female finches preferred symmetrically legbanded males over asymmetrically banded ones. Since 1999, this paper has been cited more than five times as often as a partially replicated study that found no effect of legband asymmetry on female preference (Jennions 1998). Even more surprising, a conceptually replicated study by Waas and Wordsworth (1999), which agreed with the main conclusion of Swaddle and Cuthill (1994), has only been cited twice. Also of note, Waas and Wordsworth (1999) did not cite Jennions (1998). This example raises two points. First, contradictory evidence is at times ignored. Second, unresolved conclusions need more study, particularly when the experiment is easily performed, the model organism is readily available, and the work is relatively inexpensive.

REPLICATION BATTERY

If a study is replicated, it is generally not replicated often. This presents a serious dilemma if the replicate fails to support the original study because we do not know whether the experiment was replicated improperly or the original result was in error. To counter such a dilemma, Rosenthal (1991b) suggests using a "replication battery" (see also "systematic replication;" Hendrick 1991). The simplest type of replication battery is two replications: one as exact as possible and the other moderately dissimilar to the original study. For example, if an effect size of $\delta = 0.6$ was obtained in a study, and both replications in a battery each gave $\delta = 0.50$, one would have high confidence in the original result given its robustness in the face of moderate procedural variation (Rosenthal 1991b). If, on the other hand, neither replication produced a result consistent with the original study (e.g., $\delta = 0.1$ and $\delta = -0.1$), one would have less confidence in the original study with or without procedural variation (Rosenthal 1991b). If the more exact replication showed a similar effect size (e.g., 0.55) while the other did not (e.g., 0.1), then the original result would be reliable, but seemingly not generalizable to studies involving different procedures. A more complicated replication battery would involve several replicates of the original study with a continuum of procedural dissimilarity. Homogeneous outcomes (e.g., effect sizes) of the replicates would suggest robust results, while effect sizes that vary with the degree of dissimilarity would suggest systematic sensitivity to procedural variations (Rosenthal 1991b).

DO *P*-VALUES INDICATE REPLICABILITY?

Over the past decade, considerable debate within the social sciences community has focused on the (in)significance of null hypothesis significance tests (NHST) (e.g., Harlow et al. 1997; Killeen 2005; Sohlberg and Andersson 2005). This debate is beginning to occur among evolutionists and ecologists (e.g., Johnson 1999; Stoehr 1999; Mogie 2004). At the center of the controversy is what *p*-values do and (most importantly) do not tell us about research findings. The *p*-value provided by a null hypothesis significance test gives only a measure of the probability of obtaining a result as extreme as (or more extreme than) the observed data under the null hypothesis. Less clear is the notion that *p*-values indicate the probability that a result would be obtained upon replication of the study (Shaver 1993). Critics have argued that *p*-values do not provide any confidence in the replicability of research outcomes (e.g., Carver 1978; Shaver 1993), and others argue they do (Melton 1962; Oakes 1986; Rosnow and Rosenthal 1989; Greenwald et al. 1996; Killeen 2005).

Greenwald et al. (1996) point out that those opposed to interpreting *p*-values as confidence in the replicability of research outcomes were interpreting replicability in an estimation context (i.e., direction and size of effect). In that context, opponents are correct. Arguing from an "estimation" point of view, Shaver (1993) states that statistical significance provides no information about the probability that replications of a study would yield the same result. He based his argument on Rosenthal's (1991b) conclusion that, if sample varies among the studies, different results (e.g., different *r* values) could yield similar *p*-values, while identical results (e.g., same *r* values) could yield different *p*-values. Therefore, the question of whether any or all outcomes of replicated studies are statistically significant is irrelevant (Shaver 1993). What is important "is whether an effect size of a magnitude judged to be important has been consistently obtained across valid replications" (Shaver 1993:304; see also Rosenthal 1991b).

Conversely, in a NHST context, *p*-values do provide a measure of confidence in the replicability of null hypothesis rejection, unlike that for an effect size or confidence interval (Greenwald et al. 1996). This idea was advanced by several workers (Melton 1962; Oakes 1986; Rosnow and Rosenthal 1989) despite a lack of logical support or theoretical backing (Krueger 2001). Replicability is defined by Greenwald et al. (1996) as the estimated probability that an exact replication of an initial null hypothesis rejection will similarly reject the null hypothesis. Also, it is intended only in its NHST context of repeating the dichotomous "reject/fail to reject" conclusion and not in its estimation context of proximity between point or interval estimates.

In the NHST context, replicability can be calculated as the power of an exact replication study which is approximated by the formula,

$$\text{Replicability} = 1 - \mathrm{P}\left(z \leq \frac{t_{crit} - t_1}{\sqrt{1 + \dfrac{t_{crit}^2}{2 \times df}}} \right) \quad (1)$$

where t_{crit} is the critical t value required to reject the null hypothesis, df is the degrees of freedom, P() is the probability under the cumulative normal distribution, and t_1 is the observed t value from the initial study (see Greenwald et al. 1996). Thus, for two studies with the same sample size, one has greater power to replicate an experiment if the original statistical test yielded a small p-value (e.g., $\sim$0.005).

If, for example, an isolated finding had a p-value of 0.05 (or $\sim$0.005), a researcher would at most have a 50% (or $\sim$85%) chance of getting a similar null hypothesis rejection from an exact replication (Figure 1). Given the overall trend in behavioral ecology and animal behavior research of low sample sizes, and hence low statistical power (Jennions and Møller 2003), failure to replicate a $p < 0.05$ result with two studies is not surprising. Greenwald et al. (1996) stress that a $p \sim 0.05$ is an interesting but unconvincing result for an isolated study, whereas $p \sim 0.005$ is a better indicator of demonstrability. They strongly caution that their estimates of replicability are valid only for exact replications (which are practically impossible) and, due to selective or inaccurate reporting of p-values by investigators (e.g., not Bonferroni-correcting alpha-level or reporting only statistically significant tests; Nakagawa 2004), will likely be much too high in many cases. Despite the utility of p-values providing some measure of confidence in replicability, their interpretation is not a substitute for replication studies. Greenwald et al. (1996) warn that we should have less confidence in phenomena found in only a single study than in those found repeatedly in replicate studies—preferably exact replications.

Several data analysis techniques are argued to be equivalent to replication but, as with the interpretation of p-values, they simply do not have the explanatory power of planned replications (Shaver 1993). First, meta-analyses, or the post-hoc collection of studies, typically involve the analysis of studies that have few, if any, planned connections and many unknown differences among them (Shaver 1993). This does not diminish the importance of meta-analysis; it is a necessary tool for detecting broader trends, but we must keep in mind that they are only as informative as the quality of studies going into the analysis (Jennions and Møller 2003). Second, despite the claims of Thompson (1993; see also Daniel 1998; Nix and Barnette 1998), techniques such as cross-validation, bootstrapping, and jackknifing neither indicate the likelihood of replication, nor do they provide an estimate of replicability. Instead, these procedures internally replicate (Thompson 1996) and only identify the robustness of the single study's conclusions (Levin 1998). These techniques certainly have their place in data analysis, but they cannot replace external replication (Thompson 1996) or independently conducted studies (i.e., a study conducted at different sites at different times with different specific participants and operations) (Levin 1998). The question of replicability can be settled only by replication itself. Echoing the sentiments of Shaver (1993), Levin (1998) proposes that "instead of measuring the quality of research by the level of significance, it would be better judged by its consistency of results in repeated experiments [and] if a researcher

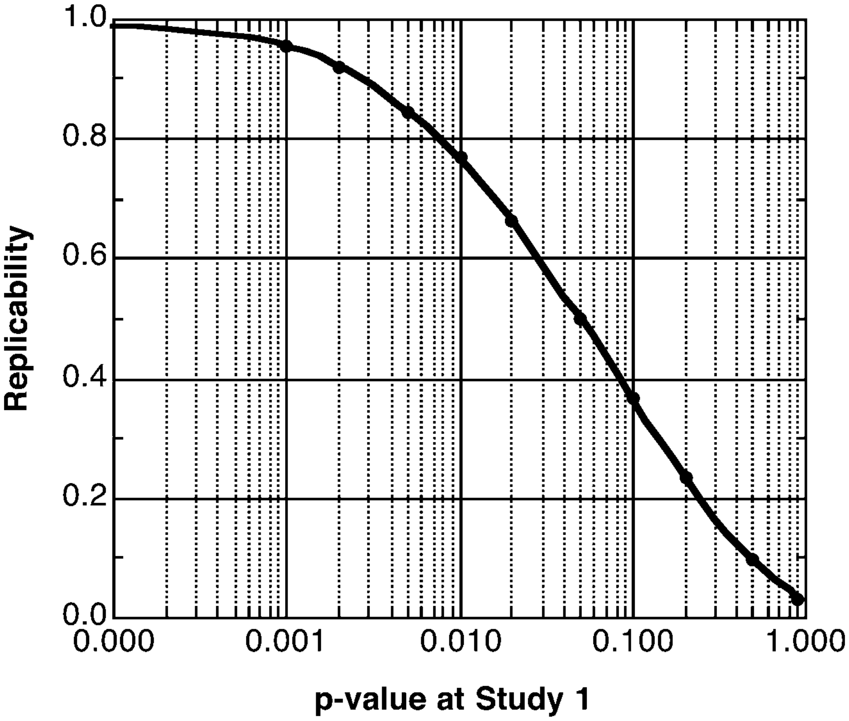

FIGURE 1. REPLICABILITY AND p-VALUES

Estimated replicability as a function of p-value (log scale) for a pair of studies each with n = 20 (df =19) and $t_{\text{crit}} = 2.093$ ($\alpha = 0.05$). Replicability is calculated using Equation 1. The region below Replicability = 0.5 are not proper replicabilities (in a null hypothesis testing context) because the outcome of the initial study did not reject the null hypothesis. This region shows the probability of a second study obtaining a null hypothesis rejection at the $\alpha = 0.05$ level when the first study did not. (Figure modified from Figure 1c in Greenwald et al. 1996)

does obtain the same result . . . more than once, it strengthens the conclusion that the results are not due to chance" (p 92).

DO BEHAVIORAL ECOLOGISTS REPLICATE?

The logical first step in addressing this issue is to ask whether we replicate experimental work and, if so, determine what type of replications we usually perform. To get an idea of how behavioral ecologists replicate studies, I categorized empirical studies (theoretical papers and reviews were excluded) published in one volume of each of three top-ranked (2005 ISI Journal Citation Reports®, The Thompson Corporation) journals in the field of behavioral ecology and animal behavior: *Behavioral Ecology and Sociobiology* (Volume 56), *Animal Behaviour* (Volume 67), and *Behavioral Ecology* (Volume 15) (Table 1). Studies were first categorized as either quasireplicates (i.e., studying same questions in different species) or true replicates (i.e., studying same questions in same species), with the latter further classified as exact (e.g., everything in-replicate study is the same as in the original study) or partial/conceptual (e.g., test same hypothesis, but population, methods, procedure, and data analysis are different) replicates. I relied on the information given by the authors in the text to categorize papers. That is, authors were typically clear in either their Introduction or Discussion as to whether the main hypothesis under investigation in the present study was tested previously in another species (i.e., quasireplication) or whether they were retesting hypotheses in the

TABLE 1

Number (percentage of volume total) of empirical research studies published categorized by type of replication

Journal	True replication		Quasireplication
	Exact	Partial and Conceptual	
Behavioral Ecology and Sociobiology	0	24 (33.8)	47 (66.2)
Animal Behaviour	0	28 (24.8)	85 (75.2)
Behavioral Ecology	0	31 (26.3)	87 (73.3)

same species, and how the present study differed from previous ones if at all (exact versus partial/conceptual replication). The distinction between partial and conceptual replications was often too difficult to determine from the information given in the paper; therefore, for simplicity, I combined these two types of replications into a single group.

Most studies published in the three journals I examined were quasireplicated (Table 1). This should be expected if we wish to find general patterns in biological phenomena among study systems. The selected journals appear to do an admirable job publishing partial/conceptual replications, as this type of study represented approximately a third (*Behavioral Ecology and Sociobiology*) to a quarter (*Animal Behaviour*) of the articles published. This may simply reflect a taxonomic bias because, for example, there appear to be many conceptually replicated studies on blue tits, pied flycatchers, red winged blackbirds, and house sparrows but few replicated studies on tropical rove beetles or marine isopods. Each journal performed poorly when it came to publishing exact replications; no exact replications were published in any of the journals examined. Is this poor representation due to exact replications not being submitted for publication or, if submitted, were they all rejected? Only journal editors can answer this query.

Why Should Behavioral Ecologists Replicate Empirical Studies?

There are several reasons why behavioral ecologists—and all biologists for that matter—should replicate empirical studies. Study replication will allow us to develop better mathematical models of biological phenomena, acquire a more accurate picture of within-species relationships, and guard against chance observations, self-deception, or deliberate falsification. I discuss each of these below.

Mathematical models allow us to understand complex biological phenomena by simplifying patterns and process (Gotelli 1998). Models are, however, only as informative as the assumptions upon which they are based (Gotelli 1998). If we aim to build models that accurately predict biological phenomena, we need to incorporate the best possible data into their parameters. Students of sperm competition have certainly benefited from the use of mathematical models (Parker 1998; Simmons 2001). A long-standing assumption in evolutionary ecology has been that sperm are cheap to produce, however, this issue is poorly understood (Wedell et al. 2002). Although evidence is beginning to emerge showing that sperm production may be costly (Nakatsuru and Kramer 1982; Olsson et al. 1997; Preston et al. 2001), the supporting studies have only been quasireplicated and are suggestive at best.

If we are to understand the precise implications of costly sperm production on male ejaculation strategies, we must have definitive evidence showing the energetic expense of sperm production. We need exact replication of these important studies.

Trying to generalize and predict biological phenomena through theoretical models or meta-analyses becomes ever more complicated when we recognize the subtleties and complexities of relationships within a single species. It is well known that an individual in a given population can behave differently at different times of the year because ecological and/or environmental variables change temporally (Foster and Endler 1999). Take the red deer (*Cervus elaphus*) on the Scottish island of Rum as an example. In the early 1970s, Trivers and Willard (1973) predicted that nonhuman, female mammals in superior condition should produce sons, whereas females in poor condition should produce daughters because males are costly to rear but provide greater fitness returns. Support for this hypothesis had been notoriously inconsistent (Clutton-Brock and Iason 1986), although some promising but inconclusive data had been collected (Clutton-Brock et al. 1984, 1986; Cockburn 1999). It was not until Kruuk and colleagues (1999) examined the deer in this population under different ecological conditions (higher population density) that the Trivers-Willard hypothesis was firmly supported: higher population density favors the production of daughters by high-ranking females because maternal resources are in short supply, and sons require more resources than daughters. If this partially replicated study had not been conducted, a major gap in our knowledge would still exist.

This situation is compounded when we examine the behavioral ecology of different populations of the same species. An example of among-population variation in behavior involves a test of the hypothesis explaining the benefits of polyandry to females. For years, behavioral ecologists pondered why females mate with more than one male when a single mating can supply enough sperm to fertilize all of a female's eggs (e.g. Simmons 2005). One hypothesis suggests that females gain genetic benefits in the form of increased offspring viability. The first empirical evidence supporting this hypothesis was published by Madsen et al. (1992). They found that in a small Swedish population of the European adder (*Vipera berus*) the number of mates was negatively correlated with the number of offspring dead at birth. In a conceptually replicated study, however, Capula and Luiselli (1994) found that in a large Italian population of *Viper berus,* females rarely mated multiply, and the number of mates did not reduce the proportion of stillborn offspring. The salient point here is that because there presumably was an increased probability of matings being between relatives in the population of adders studied by Madsen et al. (1992), their results do not apply to all populations of adder. Would Madsen et al. (1992) get the same result today if they replicated their study exactly?

Another reason for exactly replicating studies is that a chance result could become highly influential. For example, Bateman's (1948) classic work showing a fitness advantage to males (but not to females) of multiple mating has become the foundation for a significant part of sexual selection theory, and yet it has never been replicated. Not only could Bateman's results be due to random mating (Sutherland 1985), but two-thirds of Bateman's data—showing multiple mating can increase female fertility—are in conflict with his main conclusion. Surprisingly, this latter aspect of Bateman's study is seldom acknowledged (Tang-Martinez and Ryder 2005). The discrepancy between the two results may be due to Bateman's inconsistent experimental protocol (Arnold and Duvall 1994; Birkhead 2000; Tang-Martinez and Ryder 2005). Clearly, a study of such importance must be done at least twice. Another influential result that remains unreplicated is Møller's (1992) study of how male tail symmetry affects female mate preference in the barn swallow (*Hirundo rustica*). This paper effectively launched asymmetry as a potential target of mate choice. Since 1992, many investigators have devoted much research effort to testing the phenomenon in several species (i.e., quasireplication) with equivocal results—as Swaddle (2003:188) comments, " . . . the jury is still out on whether fluctuating asymmetry plays a role in sexual selection." For example, some meta-analyses have supported the hypothesis that the asymmetry of secondary sexual characters in males is an important feature of mate choice by females (e.g., Møller and Thornhill 1998; Thornhill and Møller 1998; Thornhill et al. 1999; Møller and Cuervo 2003); whereas other meta-analyses and literature reviews have called the generality of this phenomenon into question (e.g., Palmer 1999; Simmons et al. 1999; Swaddle 2003; Tomkins and Simmons 2003). Because our initial enthusiasm for new ideas can apparently sway our evaluation of scientific studies (Simmons et al. 1999; Poulin 2000;

Tomkins and Simmons 2003), paradigm-shifting research must be double-checked with exact replicate studies (Hendrick 1991; Shaver 1993). Consequently, although Møller's (1992) findings may well be replicable, a study of this importance must be replicated, preferably by investigators other than the original author(s) (sensu Rosenthal 1991b).

In addition to guarding against putting too much emphasis on a chance significant result, Branscomb (1985; see also Hooper 2002) suggests that replication helps identify scientists who were self-deceived or who deliberately falsified their work. The number ofscientists falling in the latter category is extremely rare, however, those in the former are likely to be more common than we think (Branscomb 1985; Hooper 2002), as natural selection may well have favored self-deception in humans to more convincingly deceive others (Trivers 1985). The tremendous pressure to publish research in high-impact journals requires statistically significant results supporting an important hypothesis or prediction (Koricheva 2003). This pressure may cause scientists to cease data collection prematurely when a significant result is obtained (Branscomb 1985), misinterpret the actions of their study animal (e.g., Marsh and Hanlon 2004), or bias observations in favor of their hypothesis (Rosenthal and Rubin 1978; Hooper 2002). If a breakthrough is made in a particular field, we must be compelled to check the results because, as scientists, "we have a responsibility to test evolutionary hypotheses as rigorously, as carefully and as honestly as possible" (Birkhead 2002:15). On the practical side, replication would hopefully stop others from conducting experiments which build upon the original (flawed) finding and, thus, not waste valuable time and ?resources.

The Practice of Replication

which studies should be replicated?

Ideally, scientists should replicate all experiments but give priority to replicating highly influential studies that act as foundations to particular topics of study (Hendrick 1991; Shaver 1993). However, it may require several years before we know the impact of a study on a particular field or topic of research. For example, it was not until Trivers (1972) cited Bateman's (1948) work that the latter's contribution to factors that control sexual differences was fully appreciated. Behavioral ecology is a relatively young field and burgeoning with many new subdisciplines (e.g., immunocompetence tradeoffs, cryptic female choice, strategic ejaculation, sexual conflict), and we cannot know today what will be foundational tomorrow. Notwithstanding the general importance of replication, we must concede that not all studies are worth replicating; replication will not make a trivial study worthwhile (Shaver 1993). Journal editors are left with the task of specifying criteria by which to judge those initial studies that are important and when a sufficient number of independent replications have been published to establish the reliability of a finding (Shaver 1993).

Perhaps publication in *Nature* or *Science* represents the best method of identifying an influential study. Another recommendation to facilitate the recognition of an important study is for authors to submit manuscripts to editors in a research proposal style, that is, without Results and Discussion sections, so that the publication decision is focused on study rationale and design quality (Kupfersmid 1988). Kupfersmid (1988) argues that this approach would increase turnaround time because the number of pages that need to be reviewed to make an editorial decision are reduced. Authors would save time because they would not have to analyze their data or write a discussion section until notified of publication acceptance, and would know the fate of their manuscript before data analysis. Consequently, they would not have to analyze their data in several different ways to produce results that appear significant, and it would reduce the number of irrelevant and methodologically-flawed studies in print. Kupfersmid (1988) suggests that his model of manuscript review would ultimately increase the publication of higher quality papers. This suggestion may not be practical, however, for some experimental approaches, such as multivariate designs, or in cases where models are built and explanations are derived post-hoc.

A third recommendation is that replications and multiple-experiment "packages" should comprise the basic publishable unit (Levin 1998). Rosenthal (1991b) developed a "replication index" to impartially assess replicate studies, and in particular to combat bias in "correlated replications" (see below). Briefly, this system weighs replicated studies by scoring each according to certain factors, such as time since original experiment, physical distance between investigators, personal attributes of investigators, investigators' predictions, and investigators' degree of personal contact with each other. This index ultimately provides a summary of how well-studied a given topic is irrespective of how well the outcomes of the replications match the original, as well as indicating what degree of confidence we can have in the given relationship (Rosenthal 1991b). Instead of qualitatively assessing a suite of studies on the same topic (e.g., seven "pro" and seven "con") we can tally the indices for each of the two categories and compare those sums. Therefore, replication effort would be focused on topics or relationships with either low indices or similar indices for contradictory results. Mate choice copying in guppies (*Poecilia reticulata*) would certainly benefit from such an analysis as considerable evidence from one cohort of authors suggests it exists in the wild, but studies from another group of academically unrelated authors using pet shop or feral populations suggest otherwise (Brooks 1998 and references therein).

WHEN HAS A STUDY BEEN SUCCESSFULLY REPLICATED?

Successful replication typically means a null hypothesis rejected once is rejected again in the same direction in a new study (Rosenthal 1991b). Two problems with this approach, however, are the unnecessary focus on significance level and the evaluation of replication success based on the resultant dichotomy (e.g., whether both studies rejected the null hypothesis) (Rosenthal 1991b). Several authors (e.g., Rosenthal 1991b; Shaver 1993; Greemwald et al. 1996) recommend more frequent reporting of effect size (e.g., Pearson's r) in publications because it is a summary statistic that can be used to evaluate a replication's success in a continuous fashion, such that a "degree of failure to replicate" can be stated. When investigators fail to reject the null hypothesis when replicating a study that originally rejected the null, they will erroneously claim "failure to replicate" (Rosenthal 1991b). For example, Table 2 shows a hypothetical example of the outcomes of three original studies and their replicates. The success of replication is quantified using Cohen's q, the absolute difference in effect sizes given as Fisher's z transformation of r (Rosenthal 1991b). In the first set of studies (A), examination of the p-values and effect sizes shows a clear failure to replicate (large q-value). In the second set of studies (B), the equivalent p-values and effect sizes are simply a function of sample size; replication set (B) shows more successful replication than study set (A). The final set of studies (C) shows that, despite the replicate study having a nonsignificant p-value, compared with the original study, their effect size estimates are identical, and thus, the replicate study was in fact successful (small q-value). This approach evaluates the success of a replication in a continuous fashion and states the degree to which a study has been successfully replicated, not whether it *was* or *was not* successful (Rosenthal 1991b). Alternatively, because Cohen's q is distributed as Z, the standard normal deviate, we can test whether two values of q differ statistically (Rosenthal 1991b).

As the number of replications in a study set increases, meta-analytic techniques will be required to determine replication success (Rosenthal 1991a,b). The statistical heterogeneity of effect size estimates is based on the Z or χ^2 distribution for two-study and three- or more-study situations, respectively (Rosenthal 1991a).

A common misconception is that if an effect is "real," it should be found statistically significant upon replication (Rosenthal 1991b). However, given the low power of many studies in behavioral ecology and animal behavior (Jennions and Møller 2003), this is an unreasonable expectation. Rosenthal (1991b) illustrated this fallacy with a hypothetical example. If an effect in nature is known to be $\delta = 0.50$ and an investigator studies it using 64

TABLE 2

A hypothetical comparison of three replicated studies in which conclusions are drawn using p*-values versus Cohen's* q

	Replicated studies					
	A		B		C	
	Original	**Replicate**	**Original**	**Replicate**	**Original**	**Replicate**
N	34	16	95	17	102	31
p (two-tailed)	0.0002	0.44	0.05	0.05	0.0098	0.20
Z(p)	3.55	-0.77	1.96	1.96	2.33	1.28
[a]r	0.61	-0.19	0.20	0.48	0.23	0.23
Z(r)	0.71	-0.19	0.20	0.52	0.23	0.23
Cohen's q	0.9		0.32		0.00	
p conclusion	failure to replicate		successful replication		failure to replicate	
q conclusion	failure to replicate		failure to replicate		successful replication	

[a] calculate $r = \frac{Z(p)}{\sqrt{N}}$ and use to look up $Z_{(r)}$ in Fisher's *z* transformation for correlation coefficients, *r*, statistical table (e.g., Table B.18, Zar 1999).

subjects with power equalling 50%, there is a one-in-four chance of both studies obtaining statistical significance at the 0.05 level (i.e., 25% of both studies have *p*-value > 0.05), and a one-in-eight chance of two replications (i.e., half as likely as one successful replication, 25% x 0.5) being statistically significant. Hence, because of the low power of many studies in evolution and ecology (Jennions and Møller 2003), there is no reason to expect a high proportion of significant results even if the effect in nature is real and important. I note that a low-power replicate study achieving a similar effect size as the original study is not necessarily a successful replication in an estimation context. This is because low power not only means a lack of statistical significance, but it can also yield imprecise parameter estimates.

In a further attempt to ascertain how well studies in animal behavior and behavioral ecology are replicated, I calculated the "failure to replicate" (Rosenthal 1991b). To do this, I haphazardly selected a single partially/conceptually replicated study from each journal volume examined above (Do Behavioral Ecologists Replicate?) as well as the original study cited by the authors that examined the same hypothesis. The failure to replicate is quantified as Cohen's *q* and is simply the difference between the Z(r) values for each study (Table 3); larger values of Cohen's *q* indicate greater differences in effect size (Rosenthal, 1991b). Table 3 shows that the studies on juvenile perch (*Perca flavescens*) exhibited the poorest replication (q = 1.576) despite both studies having *p*-values < 0.001, whereas the work on the collard flycatchers (*Ficedula albicollis*) showed the best replication (q = 0.29) despite having very dissimilar *p*-values. These examples again highlight the importance of examining effect sizes versus significance values. I suggest that Cohen's *q* (or heterogeneity of effect sizes for three or more studies, Rosenthal 1991b) should be a required part of a paper whenever a relationship is retested in the same species. This would provide readers with a better ability to judge the importance of relationships rather than rely on "but see" or take the author's word that their examination is similar to other studies.

WHY ARE STUDIES SELDOM REPLICATED?

There are obviously limits to what we can replicate as evolutionists and ecologists; we are not able to easily replicate studies on gorilla (*Gorilla gorilla*) behavior in the Congo. However, most of our theories are developed and tested with model organisms (e.g., *Drosophila,* guppies, zebra finches) that are, not by coincidence, much more amenable to replication. So, why are studies seldom replicated? The answer may be more sociological than practical.

TABLE 3
The degree of failure to replicate among three sets of studies

	Relationship between competitive ability of juvenile perch (*Perca flavescens*) and food intake		**Relationship between the size of a male collared flycatcher's (*Ficedula albicollis*) forehead patch area and the sex ratio of his brood**		**Relationship between parental sex and time spent provisioning larvae in the burying beetle (*Nicrophorus vespilloides*)**	
	Staffan et al. (2002)	Westerberg et al. (2004)	Ellegren et al. (1996)	Rosivall et al. (2004)	Bartlett (1988)	Smiseth and Moore (2004)
N	40	48	79	57	219	62
p (two-tailed)	<0.001	<0.001	<0.01	0.933	>0.05	<0.001
r	0.9884	0.747	0.2942	0.0132	0.0289	0.859
Z(r)	2.572	0.996	0.3032	0.0132	0.0289	1.288
Cohen's q ($Z_{r1} - Z_{r2}$)	1.576		0.29		1.259	

The studies published in 2004 were chosen haphazardly from those studies considered partial/conceptual replications from *Animal Behaviour, Behavioral Ecology and Sociobiology,* and *Behavioral Ecology,* respectively

The replication of empirical research does not seem to be part of the culture in evolution and ecology (Palmer 2000; Birkhead 2002), as it is in chemistry, physics (Hendrick 1991), or even medicine or molecular biology (Palmer 2000). Given the many factors working against replication in evolution and ecology, this attitude is unlikely to change soon. There are many reasons why we so rarely replicate studies including, for example, a dearth of adequate and pertinent information in the original study to allow replication (Thompson 1996) or a lack of stamina, time, and funds to conduct one's studies at least twice (Greenwald et al. 1996; Thompson 1996). Perhaps because a phenomenon such as animal behavior is complex and labile and, in field studies, environmental conditions vary in both space and time, replication is considered too difficult (Palmer 2000). However, "this shouldn't stop us" (Birkhead 2002:15) because this is " . . . precisely why replication is valuable in the first place—to judge just how repeatable a result is" (Palmer 2000:474).

Most importantly, the paucity of replication is likely because of a general disdain by thesis dissertation committees (Nickerson 2000) and journal editors for nonoriginal research (Kupfersmid 1988; Neuliep and Crandall 1991; Nix and Barnette 1998; Palmer 2000; Birkhead 2002; DeCoursey 2006); however, exceptions do exist (Palmer 2000). Perhaps graduate students (or more accurately their dissertation committees) should make study replication at least one part of their thesis (Palmer 2000). If a student's thesis extends previously published empirical research then, for example, one chapter of their thesis could replicate the previous salient research. However, the research completed by graduate students will often directly affect the future funding of the principal investigator under which they work. Therefore, given that funding agencies and grant-awarding bodies favor novel research, why would the grant holder (i.e., thesis supervisor and principal investigator among others) devote their valuable time and resources to replicate work that will likely be rejected as "unoriginal" or relegated to a low-ranked journal, if published at all?

Is it possible that the identity of the researcher influences the likelihood that a replicated study is published? If the replication is conducted by an unknown graduate student, we may be less inclined to accept the findings than if it was conducted by an established researcher with a reputation as a careful investigator. In contrast, if the replication was conducted by an investigator for whom some doubt existed about their research integrity, we might not accept it. Perhaps this problem could be alleviated by instituting a double-blind manuscript review process in more of our journals.

The lack of replication in the fields of evolution and ecology could be improved to some extent by journal editors and referees looking more positively upon replicated studies and being more sympathetic to those retesting fundamental ideas in ecology and evolution (Kupfersmid 1988; Neuliep and Crandall 1991; Palmer 2000; Birkhead 2002). Shaver (1993) argues that journal editors should not only encourage the reporting of replications, but in many instances demand replication before results can be published (e.g., Maddox et al. 1988). Critics of this approach suggest that editors confuse replication with reviews of research (Shaver 1993). Also, there is an implied bias when a researcher replicates their own work, a phenomenon termed "correlated replication" (Rosenthal 1991b; Nix and Barnette 1998). Rosenthal's (1991b) replication index would help to alleviate this bias.

A shift in editorial policy toward favoring replicated studies should be easier now than it was years ago. Historically, critics argued that replicated studies take up valuable journal space that could be used instead for novel findings (Shaver 1993). Today, however, many journals are online and may have considerable space available for online publication. Therefore, I recommend that journals publish replicated studies as part of their online editions. In the long run, the outcome should be better studies, an increased number of replicated studies, and greater knowledge of productivity (Shaver 1993).

Conclusion

Animal behaviorists and behavioral ecologists need to engage in more study replication of all types (Hendrick 1991; Shaver 1993; Levin 1998; Nickerson 2000; Palmer 2000). However, in order to maximize the potential value of such an endeavor, we must proceed with sound judgement. Following Hendrick's (1991) recommendations, conceptual and partial replications should be performed only if one is willing to accept a high risk of failure and only when the initial study is important and procedural changes considerably reduce labor investment. Similarly, exact replications should be reserved for important or influential (i.e., foundational) studies, as this type of replication can be expensive, time-intensive, and tedious. If a study is worthy of exact replication, then the researcher should consider systematic replication or a replication battery.

Animal behaviorists and behavioral ecologists need to follow the lead of the psychological and social sciences and engage in open and critical debate about issues surrounding the replication of empirical research. More practically, we need to report effect sizes and confidence intervals in addition to *p*-values in our research papers to facilitate comparisons between the original and replicate studies.

In general, we must realize that successful replication of a research outcome does not guarantee that the next attempt at replication will be successful as well (Nickerson 2000). Replication simply offers a justifiable increase in confidence that further replication is possible, while simultaneously supporting the theories that predicted the outcome (Nickerson 2000).

ACKNOWLEDGMENTS

I thank Darryl Gwynne, Gary Sprules, Kevin Judge, Michael Jennions, Andrew Cockburn, Rich Palmer, and two anonymous referees for insightful discussion and comments on the manuscript. This research was supported by a National Science Engineering and Research Council (NSERC) grant to D T Gwynne.

REFERENCES

Anonymous. 2006. Beautification and fraud. *Nature Cell Biology* 8:101-102.

Arnold S J, Duvall D. 1994. Animal mating systems: a synthesis based on selection theory. *American Naturalist* 143:317-348.

Bacon F (edited by W A Wright). 1926. *The Advancement of Learning*. Fifth Edition. Oxford: Clarendon Press.

Bartlett J. 1988. Male mating success and paternal care in *Nicrophorus vespilloides* (Coleoptera, Silphidae). *Behavioral Ecology and Sociobiology* 23:297-303.

Bateman A J. 1948. Intra-sexual selection in *Drosophila. Heredity* 2:349-368.

Birkhead T R. 2000. *Promiscuity: An Evolutionary History of Sperm Competition.* Cambridge (MA): Harvard University Press.

Birkhead T R. 2002. Of moths and men (book review). *International Society for Behavioral Ecology Newsletter* 14:15-16.

Branscomb L M. 1985. Integrity in science. *American Scientist* 73:421-423.

Brooks R. 1998. The importance of mate copying and cultural inheritance of mating preferences. *Trends in Ecology & Evolution* 13:45-46.

Capula M, Luiselli L. 1994. Can female adders multiply? *Nature* 369:528-528.

Carver R P. 1978. The case against statistical significance testing. *Harvard Educational Review* 48:378-399.

Clutton-Brock T H, Albon S D, Guinness F E. 1984. Maternal dominance, breeding success and birth sex ratios in red deer. *Nature* 308:358-360.

Clutton-Brock T H, Albon S D, Guinness F E. 1986. Great expectations: dominance, breeding success and offspring sex ratios in red deer. *Animal Behaviour* 34:460-471.

Clutton-Brock T H, Iason G R. 1986. Sex ratio variation in mammals. *Quarterly Review of Biology* 61:339-374.

Cockburn A. 1999. Deer destiny determined by density. *Nature* 399:407-408.

Colegrave N, Ruxton G D. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology* 14:446-447.

Coyne J A. 1998. Not black and white (book review). *Nature* 396:35-36.

Daniel L G. 1998. Statistical significance testing: a historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools* 5:23-32.

DeCoursey T E. 2006. It's difficult to publish contradictory findings. *Nature* 439:784.

Ellegren H, Gustafsson L, Sheldon B C. 1996. Sex ratio adjustment in relation to paternal attractiveness in a wild bird population. *Proceedings of the National Academy of Sciences USA* 93:11723-11728.

Fenning T M. 2004. Fraud offers big rewards for relatively little risk: we need to change the over-competitive culture that promotes publishing at all costs. *Nature* 427:393.

Fisher R A. 1974. *The Design of Experiments.* Ninth Edition. New York: Hafner Press.

Foster S A, Endler J A. 1999. G*eographic Variation in Behavior: Perspectives on Evolutionary Mechanisms.* Oxford and New York: Oxford University Press.

Gontard-Danek M-C, Møller A P. 1999. The strength of sexual selection: a meta-analysis of bird studies. *Behavioral Ecology* 10:476-486.

Gotelli N J. 1998. *A Primer of Ecology.* Second Edition. Sunderland (MA): Sinauer Associates.

Greenwald A G, Gonzalez R, Harris R J, Guthrie D. 1996. Effect sizes and *p* values: what should be reported and what should be replicated? *Psychophysiology* 33:175-183.

Harlow L L, Mulaik S A, Steiger J H. 1997. *What If There Were No Significance Tests?* Mahwah (NJ): Lawrence Erlbaum Associates.

Hendrick C. 1991. Replications, strict replications, and conceptual replications: are they important? Pages 41-49 in *Replication Research in the Social Sciences,* edited by J W Neuliep. Newbury Park (CA): Sage Publications.

Hooper J. 2002. *Of Moths and Men: An Evolutionary Tale: Intrigue, Tragedy and the Peppered Moth.* London: Fourth Estate.

Jennions M D. 1998. The effect of leg band symmetry on female-male association in zebra finches. *Animal Behaviour* 55:61-67.

Jennions M D, Møller A P. 2002. Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological Reviews* 77:211-222.

Jennions M D, Møller A P. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* 14:438-445.

Johnson D H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772.

Killeen P R. 2005. An alternative to null-hypothesis significance tests. *Psychological Science* 16:345-353.

Koricheva J. 2003. Non-significant results in ecology: a burden or a blessing in disguise? *Oikos* 102:397-401.

Krueger J. 2001. Null hypothesis significance testing: on the survival of a flawed method. *American Psychologist* 56:16-26.

Kruuk L E B, Clutton-Brock T H, Albon S D, Pemberton J M, Guinness F E. 1999. Population density affects sex ratio variation in red deer. *Nature* 399:459-461.

Kupfersmid J. 1988. Improving what is published: a model in search of an editor. *American Psychologist* 43:635-642.

Levin J R. 1998. What if there were no more bickering about statistical significance tests? *Research in the Schools* 5:43-53.

Lykken D T. 1968. Statistical significance in psychological research. *Psychological Bulletin* 70:151-159.

Maddox J, Randi J, Stewart W W. 1988. "High dilution" experiments a dilution. *Nature* 334:287-290.

Madsen T, Shine R R, Loman J, Håkansson T. 1992. Why do female adders copulate so frequently? *Nature* 355:440-441.

Marsh D M, Hanlon T J. 2004. Observer gender and observation bias in animal behaviour research: experimental tests with red-backed salamanders. *Animal Behaviour* 68:1425-1433.

Martinson B C, Anderson M S, de Vries R. 2005. Scientists behaving badly. *Nature* 435:737-738.

Melton A W. 1962. Editorial. *Journal of Experimental Psychology* 64:553-557.

Mogie M. 2004. In support of null hypothesis significance testing. *Proceedings of the Royal Society of London B* 271(Supplement):S82-S84.

Montgomerie R M, Birkhead T. 2005. A beginner's guide to scientific misconduct. *International Society for Behavioral Ecology Newsletter* 17:16-24.

Møller A P. 1992. Female swallow preference for symmetrical male sexual ornaments. *Nature* 357:238-240.

Møller A P, Cuervo J J. 2003. Asymmetry, size, and sexual selection: factors affecting heterogeneity in relationships between asymmetry and sexual selection. Pages 262-275 in *Developmental Instability: Causes and Consequences,* edited by M Polak. Oxford: Oxford University Press.

Møller A P, Jennions M D. 2001. Testing and adjusting for publication bias. *Trends in Ecology & Evolution* 16:580-586.

Møller A P, Thornhill R. 1998. Bilateral symmetry and sexual selection: a meta-analysis. *American Naturalist* 151:174-192.

Møller A P, Thornhill R, Gangestad S W. 2005. Direct and indirect tests for publication bias: asymmetry and sexual selection. *Animal Behaviour* 70:497-506.

Nakagawa S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology* 15:1044-1045.

Nakatsuru K, Kramer D L. 1982. Is sperm cheap? Limited male fertility and female choice in the lemon tetra (Pisces, Characidae). *Science* 216:753-755.

Neuliep J W, Crandall R. 1991. Editorial bias against replication research. Pages 85-90 in *Replication Research in the Social Sciences,* edited by J W Neuliep. Newbury Park (CA): Sage Publications.

Nickerson R S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* 5:241-301.

Nix T W, Barnette J J. 1998. The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *Research in the Schools* 5:3-14.

Oakes M W. 1986. *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* Chichester (UK) and New York: Wiley.

Olsson M, Madsen T, Shine R. 1997. Is sperm really so cheap? Costs of reproduction in male adders, *Vipera berus. Proceedings of the Royal Society of London B* 264:455-459.

Palmer A R. 1999. Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. *American Naturalist* 154:220-233.

Palmer A R. 2000. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics* 31:441-480.

Parker G A. 1998. Sperm competition and the evolution of ejaculates: towards a theory base. Pages 3-54 in *Sperm Competition and Sexual Selection,* edited T R Birkhead and A P Møller. San Diego (CA): Academic Press.

Pearson H. 2003. Competition in biology: it's a scoop! *Nature* 426:222-223.

Poulin R. 2000. Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society of London B* 267:787-792.

Preston B T, Stevenson I R, Pemberton J M, Wilson K. 2001. Dominant rams lose out by sperm depletion: a waning success in siring counters a ram's high score in competition for ewes. *Nature* 409:681-682.

Rosenthal R. 1991a. *Meta-analytic Procedures for Social Research.* Revised Edition. Newbury Park (CA): Sage Publications.

Rosenthal R. 1991b. Replication in behavioral research. Pages 1-29 in *Replication Research in the Social Sciences,* edited by J W Neuliep. Newbury Park (CA): Sage Publications.

Rosenthal R, Rubin D B. 1978. Interpersonal expectancy effects: the first 345 studies. *Behavioral and Brain Sciences* 3:377-415.

Rosivall B, Torok J, Hasselquist D, Bensch S. 2004. Brood sex ratio adjustment in collared flycatchers (*Ficedula albicollis*): results differ between populations. *Behavioral Ecology and Sociobiology* 56:346-351.

Rosnow R L, Rosenthal R. 1989. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist* 44:1276-1284.

Shaver J P. 1993. What statistical significance testing is, and what it is not. *Journal of Experimental Education* 61:293-316.

Simmons L W. 2001. *Sperm Competition and Its Evolutionary Consequences in the Insects.* Princeton (NJ): Princeton University Press.

Simmons L W. 2005. The evolution of polyandry: sperm competition, sperm selection, and offspring viability. *Annual Review of Ecology, Evolution and Systematics* 36:125-146.

Simmons L W, Tomkins J L, Kotiaho J S, Hunt J. 1999. Fluctuating paradigm. *Proceedings of the Royal Society of London B* 266:593-595.

Smiseth P T, Moore A J. 2004. Signalling of hunger when offspring forage by both begging and self-feeding. *Animal Behaviour* 67:1083-1088.

Sohlberg S, Andersson G. 2005. Extracting a maximum of useful information from statistical research data. *Scandinavian Journal of Psychology* 46:69-77.

Staffan F, Magnhagen C, Alanärä A. 2002. Variation in food intake within groups of juvenile perch. *Journal of Fish Biology* 60:771-774.

Stoehr A M. 1999. Are significance thresholds appropriate for the study of animal behaviour? *Animal Behaviour* 57:F22-F25.

Sutherland W J. 1985. Chance can produce a sex difference in variance in mating success and explain Bateman's data. *Animal Behaviour* 33:1349-1352.

Swaddle J P. 2003. Fluctuating asymmetry, animal behavior, and evolution. *Advances in the Study of Behavior* 32:169-205.

Swaddle J P, Cuthill I C. 1994. Preference for symmetric males by female zebra finches. *Nature* 367:165-166.

Tang-Martinez Z, Ryder T B. 2005. The problem with paradigms: Bateman's worldview as a case study. *Integrative and Comparative Biology* 45:821-830.

Thompson B. 1993. The use of statistical significance tests in research: bootstrap and other alternatives. *Journal of Experimental Education* 61:361-377.

Thompson B. 1996. AERA editorial policies regarding statistical significance tests: Three suggested reforms. *Educational Research* 25:26-30.

Thornhill R, Møller A P. 1998. The relative importance of size and asymmetry in sexual selection. *Behavioral Ecology* 9:546-551.

Thornhill R, Møller A P, Gangestad S W. 1999. The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. *American Naturalist* 154:234-241.

Tomkins J L, and Simmons L W. 2003. Fluctuating asymmetry and sexual selection: paradigm shifts, publication bias, and observer expectation. Pages 231-261 in *Developmental Instability: Causes and Consequences,* edited by M Polak. Oxford: Oxford University Press.

Trivers R. 1985. *Social Evolution.* Menlo Park (CA): Benjamin/Cummings Publishing.

Trivers R L. 1972. Parental investment and sexual selection. Pages 136-179 in *Sexual Selection and the Descent of Man, 1871-1971,* edited by B Campbell. Chicago (IL): Aldine.

Trivers R L, Willard D E. 1973. Natural selection of parental ability to vary the sex ratio of offspring. *Science* 179:90-92.

Waas J R, Wordsworth A F. 1999. Female zebra finches prefer symmetrically banded males, but only during interactive mate choice tests. *Animal Behaviour* 57:1113-1119.

Wedell N, Gage M J G, Parker G A. 2002. Sperm competition, male prudence and sperm- limited females. *Trends in Ecology & Evolution* 17:313-320.

Westerberg M, Staffan F, Magnhagen C. 2004. Influence of predation risk on individual competitive ability and growth in Eurasian perch, *Perca fluviatilis. Animal Behaviour* 67:273- 279.

Zar J H. 1999. *Biostatistical Analysis.* Fourth Edition. Upper Saddle River (NJ): Prentice Hall.