Animal Behaviour 93 (2014) 15–23

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

When righting is wrong: performance measures require rank repeatability for estimates of individual fitness

Christina M. Davy[a,b,*], James E. Paterson[c,1], Ashley E. Leifso[a,2]

[a] Wildlife Preservation Canada, Guelph, ON, Canada
[b] Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada
[c] Department of Biology, University of Ottawa, Ottawa, ON, Canada

ARTICLE INFO

Article history:
Received 3 January 2014
Initial acceptance 19 February 2014
Final acceptance 28 March 2014
Published online
MS. number: A14-00007R

Keywords:
Apalone spinifera
Emydoidea (Emys) blandingii
fitness proxy
freshwater turtle
individual repeatability
performance measure
rank correlation
rank repeatability
righting time
righting response

Fitness proxies such as performance measures are used to quantify relative fitness in systems where direct measurements are unobtainable. To provide meaningful results at the individual level, fitness proxies must demonstrate not only repeatability, as measured by high intraclass correlation coefficients, but also rank repeatability. Here we illustrate the importance of rank repeatability in fitness proxies using a commonly employed example: righting time in hatchling turtles. Our results show that individual righting time varies strongly among trials and is not replicable enough to provide repeatable rankings of individuals or clutches. To illustrate the potential implications of this finding, we use our data to test the predication that larger turtles have faster righting times, using three consecutive trials of righting time. The resulting conclusions vary substantially among trials. Thus, we conclude that righting time does not meet the criterion of rank repeatability required for estimates of relative individual fitness, performance or phenotypic quality. Researchers employing similar proxies should assess the rank repeatability of a proxy before applying it to questions of relative individual fitness. If a measure shows satisfactory repeatability, the final test for a fitness proxy is to demonstrate a correlation with actual fitness, ideally in the organism's natural habitat.

© 2014 The Association for the Study of Animal Behaviour. Published by Elsevier Ltd. All rights reserved.

Performance measures and proxies of fitness or phenotypic quality are often used to study wild populations where a direct measure of fitness is not possible. Commonly used examples include body condition (Bradford et al., 2012), stress hormone concentrations (Bonier, Martin, Moore, & Wingfield, 2009), locomotor speed (Huey & Dunham, 1987; Langkilde, Lance, & Shine, 2005), symmetry (Alford, Bradfield, & Richards, 2007; Shine, Langkilde, Wall, & Mason, 2005), and most recently, personality (Ibáñez, Marzal, López, & Martin, 2013; Jandt et al., 2013; Menzies, Timonin, McGuire, & Willis, 2013; Seyfarth, Silk, & Cheney, 2012). Proxies of fitness or 'phenotypic quality' ideally meet several criteria. There must be significant among-individual variation for the trait to be subject to selection, and of evolutionary importance (Wilson & Nussey, 2010). The trait should be relevant to the study species' ecology and should also be correlated with an individual's actual fitness (e.g. Isden, Panayi, Dingle, & Madden, 2013; Wikelski & Romero, 2003; Wilson & Nussey, 2010), although such correlations can be extremely difficult to demonstrate, especially in long-lived or rare species. Most importantly, the measured trait must provide a repeatable estimate of fitness or performance.

Repeatability is particularly critical with potentially plastic behavioural traits such as nesting phenology or locomotor performance. Repeatability of behaviour is typically measured using the intraclass correlation coefficient (ICC; Lessells & Boag, 1987), and many behaviours show high individual repeatability based on ICC (Bell, Hankison, & Lakowski, 2009; Briffa & Greenaway, 2011; Dingemanse, Kazem, Réale, & Wright, 2010; Huey & Dunham, 1987). But individual fitness is a relative quantity, not an absolute one. Therefore, for a fitness proxy or performance measure to be used to compare fitness among individuals with different traits (rather than among groups subjected to different treatments), these measures must also demonstrate rank repeatability: they must rank individuals in a consistent order when the

[*] Correspondence and present address: C. M. Davy, Natural Resources DNA Profiling & Forensic Centre, 2140 East Bank Drive, Trent University, Peterborough, ON K9J 7B8, Canada.
E-mail addresses: purple_salamander@hotmail.com, christinadavy@trentu.ca (C. M. Davy).
[1] *E-mail address*: jpate066@uottawa.ca; james.earle.paterson@gmail.com (J. E. Paterson).
[2] *E-mail address*: aleifso@gmail.com (A. E. Leifso).

http://dx.doi.org/10.1016/j.anbehav.2014.04.013
0003-3472/© 2014 The Association for the Study of Animal Behaviour. Published by Elsevier Ltd. All rights reserved.

measurement is repeated (Huey & Dunham, 1987; Laming, Jenkins, & McCarthy, 2013; Refsnider, 2013). Most studies assessing repeatability of a trait calculate the ICC, but there is a difference between high repeatability (high ICC) and rank repeatability (repeatable ranking of individuals). Rank correlation is often not assessed and its importance when considering questions related to relative individual fitness cannot be overstated.

The adaptive significance of various genetic or phenotypic traits at the individual level can be tested using fitness proxies that demonstrate rank repeatability. For example, male blue tits, *Cyanistes caeruleus*, that successfully sire offspring through extrapair copulations with a neighbouring female are typically larger than that female's social mate (Kempenaers, Verheyena, & Dhondia, 1997). Kempenaers et al. (1997) point out that this study relies on comparing the relative size and reproductive success of neighbouring males (the relative rank by size of individual males competing to fertilize a particular clutch), not on comparing size among all males within a population. In another study testing the effect of paternity on fitness, nestling bluethroats, *Luscinia svecica*, fathered by extrapair males displayed significantly higher immune responses than their half-siblings fathered by the mother's social mate (Johnsen, Andersen, Sunding, & Lifjeld, 2000). Comparison of sibling pairs within a clutch controls for maternal and nest effects and reveals effects of paternal variation that may not be detectable by comparing immune response of extrapair and within-pair offspring across an entire population.

In long-lived organisms where direct measures of fitness are challenging, repeatable performance measures or fitness proxies could also be used to test hypotheses about the relative fitness of half-siblings sired by different fathers, or the fitness of offspring produced by different pairs of mates (Banger, Blouin-Demers, Bulté, & Lougheed, 2013; Byrne & Roberts, 2000). However, the appropriate unit of measurement for such questions is the individual offspring, not the clutch. Therefore, this approach requires a measure of relative individual fitness that demonstrates rank repeatability. In this study we assess a commonly used performance measure in hatchling turtles to assess whether it demonstrates rank repeatability, and test potential effects of low rank repeatability in a performance measure on hypothesis testing.

Fitness, performance and phenotypic quality in hatchling turtles and squamates are often estimated using righting time: the time it takes an individual to right itself after being placed on its back (Burger, 1989; Freedberg, Stumpf, Ewert, & Nelson, 2004; Micheli-Campbell, Campbell, Cramp, Booth, & Franklin, 2011; Mullins & Janzen, 2006; Patterson & Blouin-Demers, 2008; Steyermark & Spotila, 2001). Righting time is considered an ecologically relevant parameter for all hatchling turtles because they risk overturning as they disperse from nests to aquatic environments, and because mortality from predators and desiccation during this dispersal is high (Burger, 1976; Finkler & Claussen, 1997). There is growing evidence that environmental conditions such as temperature and hydration during development affect righting response in hatchling turtles (Delmas, Baudry, Girondot, & Prevot-Juillard, 2007; Finkler, 1999; Freedberg et al., 2004; Micheli-Campbell et al., 2011; Mullins & Janzen, 2006; Refsnider, 2013). However, empirical evidence that righting time accurately predicts fitness is equivocal. Delmas et al. (2007) found that righting time was positively correlated with growth and survival rates in nests incubated at fluctuating temperatures, but not in a second, paired sample of nests incubated at constant temperatures. Further confounding the situation is the variation in crypsis among different species, which may also affect optimal righting strategy. Specifically, the coloration of the plastron may be cryptic enough that staying still provides a larger fitness advantage than righting and moving to safety. Finally, a clear link between fitness and righting time (whether faster or slower) has not yet been established (Delmas et al., 2007; Refsnider, 2013).

Righting time is quantified in different ways among studies. Some studies quantify active righting time, or the time spent actively trying to turn onto the plastron (e.g. Ben-Ezra, Bulté, & Blouin-Demers, 2008; Freedberg et al., 2004; Micheli-Campbell et al., 2011). Others also measure latency time: the time the turtle spends lying passively on the carapace before attempting to right itself (e.g. Delmas et al., 2007; Rasmussen & Litzgus, 2010; Refsnider, 2013). Still others measure total righting time, the time from inverting the turtle until the turtle rights itself, equal to the sum of latency time and active righting time (Finkler, 1999; Mullins & Janzen, 2006; Steyermark & Spotila, 2001). Whichever measure is used, rank repeatability based on the results remains a key requirement for a proxy of individual fitness. To our knowledge, the repeatability of relative fitness estimates based on righting time (i.e. rank repeatability) has never been tested, as studies have either measured each turtle once (Finkler, 1999; Freedberg et al., 2004; Paitz, Clairardin, Griffin, Holgersson, & Bowden, 2009; Riley & Litzgus, 2013), measured each individual several times but selected the fastest time for analysis (Banger et al., 2013; Ben-Ezra et al., 2008; Mullins & Janzen, 2006), used the average of several trials (Maulany, Booth, & Baxter, 2012), or used the clutch as the unit of measurement (Refsnider, 2013).

Here, we investigate the hypothesis that righting response can be used as an indicator of relative fitness or phenotypic quality of individuals by testing the following predictions: (1) interindividual variation in righting response is higher than intraindividual variation; (2) individual repeatability of righting time (measured by the intraclass correlation coefficient) is high; and (3) individuals within a clutch rank in a similar order in repeated trials. We also test two competing hypotheses about individual response to multiple righting trials. If hatchlings learn to right themselves more quickly with experience (or right themselves more quickly because of handling stress), then individuals should right themselves more quickly in consecutive trials. Alternatively, if hatchlings tire with multiple trials, then individuals should right themselves more slowly in consecutive trials. Figure 1 shows two extreme examples of the expected results if righting time is a replicable measure of relative performance or fitness. In these examples, individual righting time changes among trials, but the rank of individuals within a clutch is stable. If individuals show a consistent response to multiple trials, we predicted three possible responses: (1) righting time of an individual could be stable among trials, (2) it could slow as hatchings tire (righting time increases with progressive trials; Fig. 1a), or (3) hatchlings could learn to right themselves more quickly with experience (righting time decreases with progressive trials; Fig. 1b).

METHODS

Ethical Note

Animal care protocols were approved independently by the Animal Care Committee of the Royal Ontario Museum (ROM AUP number 2011-18) and the Wildlife Animal Care Committee of the Ontario Ministry of Natural Resources (OMNR) (WACC number 11-249). This research was also authorized under Wildlife Scientific Collector's Authorization 1062210 and permit AY-B-013-11 from the OMNR, and a research authorization from Ontario Parks.

Nest Collection and Care

We collected turtle nests in June and July 2011, during an ongoing conservation project in a wetland complex on Lake Erie,

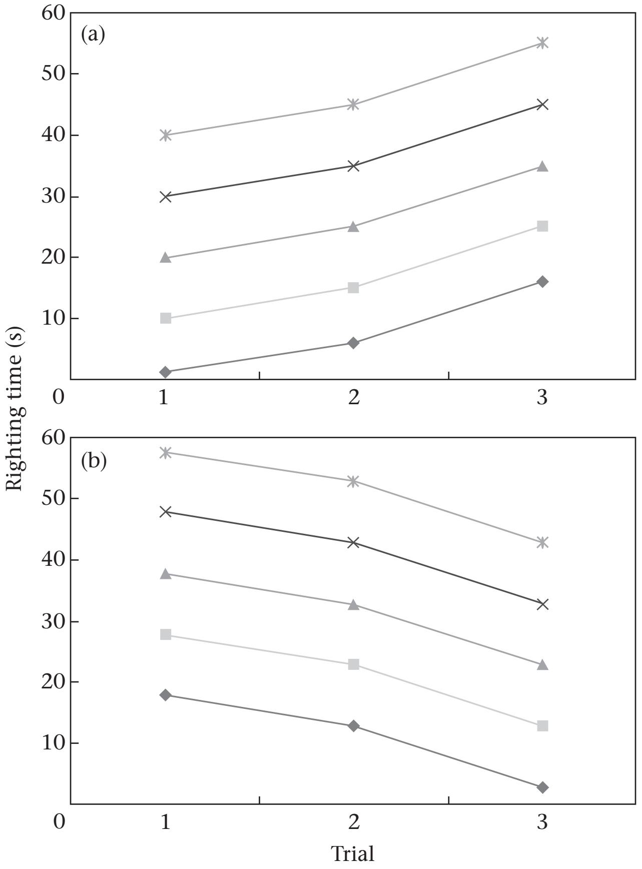

Figure 1. Hypothetical righting times of hatchling turtles in three consecutive trials if individuals demonstrate (a) exhaustion or (b) learning. Both examples assume that individuals respond to additional trials similarly and there is no interaction of trial with individual.

ON, Canada (exact location withheld to prevent increased collection pressure on populations). Predation by racoons, *Procyon lotor*, striped skunks, *Mephitis mephitis*, and Virginia opossums, *Didelphis virginiana*, at this site destroys more than 95% of the nests in some years. Nesting females were observed from a distance and were not disturbed until egg laying was complete or nesting was aborted. Females were then measured, marked and released at the nest site. All nests had different mothers. We targeted nests of the spiny softshell turtle, *Apalone spinifera*, and the Blanding's turtle, *Emydoidea blandingii*.

We incubated eggs in vermiculite mixed with water (1:1 by weight), incubating each clutch separately. Because *E. blandingii* exhibits temperature-dependent sex determination (TSD), clutches were halved and incubated at both 26 °C and 30 °C to produce both males and females. In *A. spinifera*, which does not exhibit TSD, all eggs were incubated at 30 °C. We weighed incubation containers on the day of collection and monitored their weight during incubation, adding distilled water as needed to maintain constant hydric conditions.

Hatchling Care and Release

Once clutches began to hatch we monitored the eggs daily. When a turtle emerged fully from the eggshell we placed it in a small tub with damp paper towel until the yolk sac was absorbed fully. Hatchlings with fully absorbed yolk sacs were kept in shallow water (20–30 mm deep) with their clutchmates until release. Each turtle was measured (straight line carapace length ±0.1 mm; SCL) and weighed (±0.5 g), and given three trials of righting time (see below). Finally, each turtle was marked with visible implant elastomer and a 1.5 mm decimal coded wire tag (Northwest Marine Technologies, Shaw Is., WA, U.S.A.; Davy, Coombes, Whitear, & Mackenzie, 2010) and released after 48 h of observation. Coded wire tags have been used on hatchling turtles previously without any known harm or impediment to movement (Schwartz, 1981), and no signs of distress, swelling or impaired movement have been observed in their use on over 4000 hatchlings involved in this conservation project.

Hatchling righting time trials all occurred within 1 day of yolk sac absorption, which occurred approximately 24–48 h after hatching. All hatchlings within a clutch were tested on the same day. Hatchlings were released during daylight hours at the edge of aquatic habitats adjacent to their original nest locations, and within 100 m of the original nest site.

Measuring Righting Time

We gave each hatchling three successive righting trials under identical conditions. To conduct a trial, a hatchling was placed on its carapace on a smooth table top. We recorded total righting time to the nearest second (s) with a digital stopwatch (time elapsed from the moment the hatchling was overturned until the moment it successfully righted itself). Hatchlings that righted themselves in less than 1 s were assigned a score of 0.5 s. Hatchlings that did not right themselves within 60 s were manually righted and assigned a score of 60 s. We chose to end trials at 60 s because we reasoned that the ecological relevance of this trait is restricted to a short time span (a predator that detected an overturned hatchling would have ample opportunity for predation within a minute). We also wanted to minimize the stress of trials to the hatchlings, and 60 s trials were sufficient to test the individual repeatability of righting time and the variation of individual rank among trials. We chose to conduct three trials because this was the lowest number of trials that would allow us to test whether individual righting time was repeatable while also minimizing the length of time hatchlings were in captivity and the amount of time spent processing individuals. Righting tests did not appear to be stressful for turtles and all individuals behaved normally and moved towards cover after release.

Each hatchling was rested for 10 s between trials. We ran trials consecutively with only a short interval between trials (1) to ensure that environmental conditions did not vary between trials for each hatchling and (2) to test our hypotheses regarding the potential roles of learning and tiring in determining individual performance. Ambient temperature was constant among trials, as performance is strongly temperature dependent in ectotherms (Huey & Stevenson, 1979; Kingsolver & Huey, 2008).

Quantifying Rank Repeatability

We assessed individual rank repeatability within each clutch based on total righting time, considering each species separately. We used Spearman rank correlation coefficient (r_S) to determine whether the rank orders of hatchlings within a clutch were significantly correlated between each of the three pairs of trials. This approach required two tests with each trial, so we adjusted the *P* value required for significance using sequential Bonferroni correction for multiple tests of the same data (Rice, 1989). Bonferroni correction reduces the chance of type I error but is sufficiently conservative to increase the chance of type II error, and we therefore considered results both with and without the correction.

Quantifying Individual and Clutch Repeatability

We constructed generalized linear mixed effects models to explain variation in total righting time due to sex, size, individual identity (individual repeatability), trial (learning or tiring), clutch (maternal genetic effects) and interactions between these variables. For *A. spinifera,* sex was determined by the presence (in males) or absence (in females) of complete black outlines around spots on the carapace (Greenbaum & Carr, 2001). For *E. blandingii*, we assumed all hatchlings incubated at 26 °C were males and all hatchlings incubated at 30 °C were females (Ewert & Nelson, 1991). The righting time of hatchlings was not normally distributed, and normality was not significantly increased with data transformations. Because of violations of the assumptions in normality and homoscedasticity for traditional linear regression, righting time was modelled using generalized linear mixed effects models using a Poisson distribution and a log-link function. Coefficients were fitted using the Laplace approximation (Raudenbush, Yang, & Yosef, 2000), and separate models were used for *A. spinifera* and *E. blandingii*.

The mixed effects models explaining variation in righting time were compared using Akaike's Information Criteria (AIC) to remove extra explanatory variables and avoid overfitting (Burnham & Anderson, 1998). Models with the lowest AIC scores were considered to be most supported. The fixed effect explanatory variables included sex, SCL (continuous variable in mm), and an interaction between these variables. The random effect explanatory variables included hatchling identity nested within clutch, clutch identity, trial and an interaction between trial and individual identity. Model construction started with all fixed effects and random effects, and factors were deleted using stepwise backwards model selection using log-likelihood tests. To examine repeatability, we calculated intraclass correlation coefficients (Lessells & Boag, 1987) using residual variance and the variance components from the random effect components. Generalized linear models were constructed in R (R Development Core Team, 2012) using the 'lme4' package (Bates, Maechler, & Bolker, 2011). All tests were conducted with $\alpha = 0.05$ unless stated otherwise.

Effect of Repeated Measurement on Hypothesis Testing

The relationship between righting time and body size (a potential co-correlate of phenotypic quality or fitness) is frequently tested in studies involving righting time (e.g. Micheli-Campbell et al., 2011; Refsnider, 2013). To investigate the effect of among-trial variation in righting time on biological conclusions, we tested whether different trials provided the same evidence for or against the 'bigger is better' hypothesis: that larger hatchlings have higher phenotypic quality or fitness than smaller hatchlings. We used a generalized linear model to quantify the relationship between body size and righting time for *A. spinifera* and *E. blandingii*. We tested this relationship for each trial independently to imitate scenarios where righting time was either measured only once (Freedberg et al., 2004; Riley & Litzgus, 2013), or based on a single measurement derived from multiple trials (e.g. mean, median or fastest righting time; Mullins & Janzen, 2006; Refsnider, 2013).

RESULTS

Hatching Success

Hatch success (percentage of fertilized eggs that hatched) was high for both *A. spinifera* (85%) and *E. blandingii* (92%); this is a higher hatch success than documented in either species under natural conditions (Bolton, Marshall, & Brooks, 2008; Congdon, Tinkle, Breitenbach, & van Loben Sels, 1983; Standing, Herman, & Morrison, 1999).

Rank Repeatability

We tested righting time in 374 hatchling turtles from 39 clutches. Of 1122 individual trials, 10% were assigned a value of 60 s because hatchlings did not right themselves in the allotted time. Table 1 summarizes performance of each species, and lists the frequency of significant ($\alpha = 0.05$) positive rank correlations between righting trials both before and after sequential Bonferroni correction. Individual rank varied greatly among trials within most clutches (Fig. 2a, b). Of 117 pairs of righting trials within clutches (39 clutches given three trials each), individual ranks based on righting time were significantly correlated in 34/117 pairs (29.1%) before sequential Bonferroni correction. Twenty-one pairs of trials (17.9%) were negatively correlated ($r_S = -0.049$ to -0.707), but none of the negative correlations was significant. After correction for multiple pairwise tests, ranks from only 5/117 pairs of trials (from 4/39 clutches) were significantly correlated at $\alpha = 0.05$. No clutches showed significant correlations in individual rank order among all three pairs of trials. Ranks of clutches based on mean righting time were significantly correlated in one of three pairs of trials for *A. spinifera* and in two of three pairs of trials in *E. blandingii* (Table 2, Fig. 2c, d).

Individual and Clutch Repeatability

Apalone spinifera

We used data from 222 individuals from 14 clutches for all generalized linear mixed effects models on righting time in *A. spinifera*. Twenty-one individuals were excluded because they were not tested in all three trials or were missing size measurements. The best model for righting time for *A. spinifera* hatchlings included significant fixed effects of sex, SCL, and an interaction between sex and SCL (AIC = 1427; Supplementary Fig. S1). There were significant random effects of hatchling identity, clutch identity, and an interaction between individual and trial. The coefficients, test statistics and associated *P* values of explanatory variables from the best model are presented in Table 3. Righting time was faster in males than in females and was negatively correlated with size. The interaction term between size and sex indicated that righting time decreased with size more in females than males (more negative slope in female hatchlings). Significant random effects of individual and clutch indicate variation in performance at the individual and clutch level, even when accounting for size (SCL) and sex differences. However, random effects of an interaction between trial and individual identity indicate that individuals' righting times responded differently to the three trials and that performance between trials was not always correlated (Fig. 3). The trial did not explain any variation in righting time among hatchlings in this species. The ICC was 0.165 for individual *A. spinifera* and 0.049 for clutch identity.

Emydoidea blandingii

Models for *E. blandingii* righting time used 152 individuals from 25 clutches. The best model for righting time for *E. blandingii* hatchlings (AIC = 1620) had a significant fixed effect of size, but not of sex (Supplementary Fig. S2). Larger turtles had faster righting times. The coefficients, test statistics and associated *P* values of explanatory variables from the best model are presented in Table 3. The random effects portion of the model indicate that individuals and clutches both significantly affect righting time. The random effect of trial was significant, and average righting time was shorter in later trials. There was a significant interaction effect of trial with

Table 1
Sample sizes and total righting time (mean ± SE) for hatchling turtles given three trials per individual, and number (proportion) of clutches exhibiting a significant ($\alpha = 0.05$) correlation in individual ranks

Species	No. clutches tested (mean clutch size)	Total no. individuals tested	Total righting time (mean±SE)			No. clutches with significant rank correlation		
						Before Bonferroni correction		After correction
			Trial 1	Trial 2	Trial 3	Between one or more trial pairs	Between all trial pairs	Between one or more trial pairs*
Apalone spinifera	14 (16)	222	5.33±0.86	5.60±0.92	5.67±0.89	6 (43%)	2 (14%)	3 (21%)
Emydoidea blandingii	25 (6.9)	152	23.91±1.96	18.96±1.80	16.36±1.74	12 (48%)	4 (20%)	1 (5%)
Total	39	374				18 (46%)	6 (15%)	4 (10%)

Results are shown for significant rank correlation between at least one pair of righting time trials and among all three pairs of trials, both before and after sequential Bonferroni correction for multiple pairwise comparisons.

* No clutch demonstrated significant ($\alpha = 0.05$) individual rank correlation among all three pairs of trials after sequential Bonferroni correction.

individuals on performance, indicating that individuals responded differently to the trials and that performance in one trial was not correlated with performance in another trial (Fig. 3). The ICC was 0.246 for individual *E. blandingii* and 0.147 for clutch identity.

Effect of intraindividual among-trial variation on hypothesis testing

Considering significance at $\alpha = 0.05$, intraindividual variation among trials had a strong effect on the outcome of hypothesis testing based on our tests of righting time (Table 4). For *A. spinifera,* the hypothesis that larger turtles perform better (right themselves more quickly) would be rejected based on trials 1 and 3, and accepted based on trial 2. For *E. blandingii*, the hypothesis would be rejected based on trial 1 and accepted based on trial 3. It would be tentatively rejected based on trial 2 ($P = 0.06$, a value sometimes interpreted in behavioural or ecological studies as 'approaching significance'; e.g. Parsons, Foster, & Osmond, 2012; Telemeco, Radder, Baird, & Shine, 2010).

DISCUSSION

Our results indicate that although there is an individual component to righting time (as with most behaviours), intraindividual variation in righting time is high, making distinguishing performance between individuals unreliable and difficult. Responses to multiple trials vary strongly among individuals and species, and ranks of individuals and clutches are not highly repeatable among trials. Individual rank was significantly correlated in only 4% of paired righting time trials within clutches (5/117 trials), and clutches were ranked in the same order in only half of the pairwise comparisons based on mean righting time of clutches.

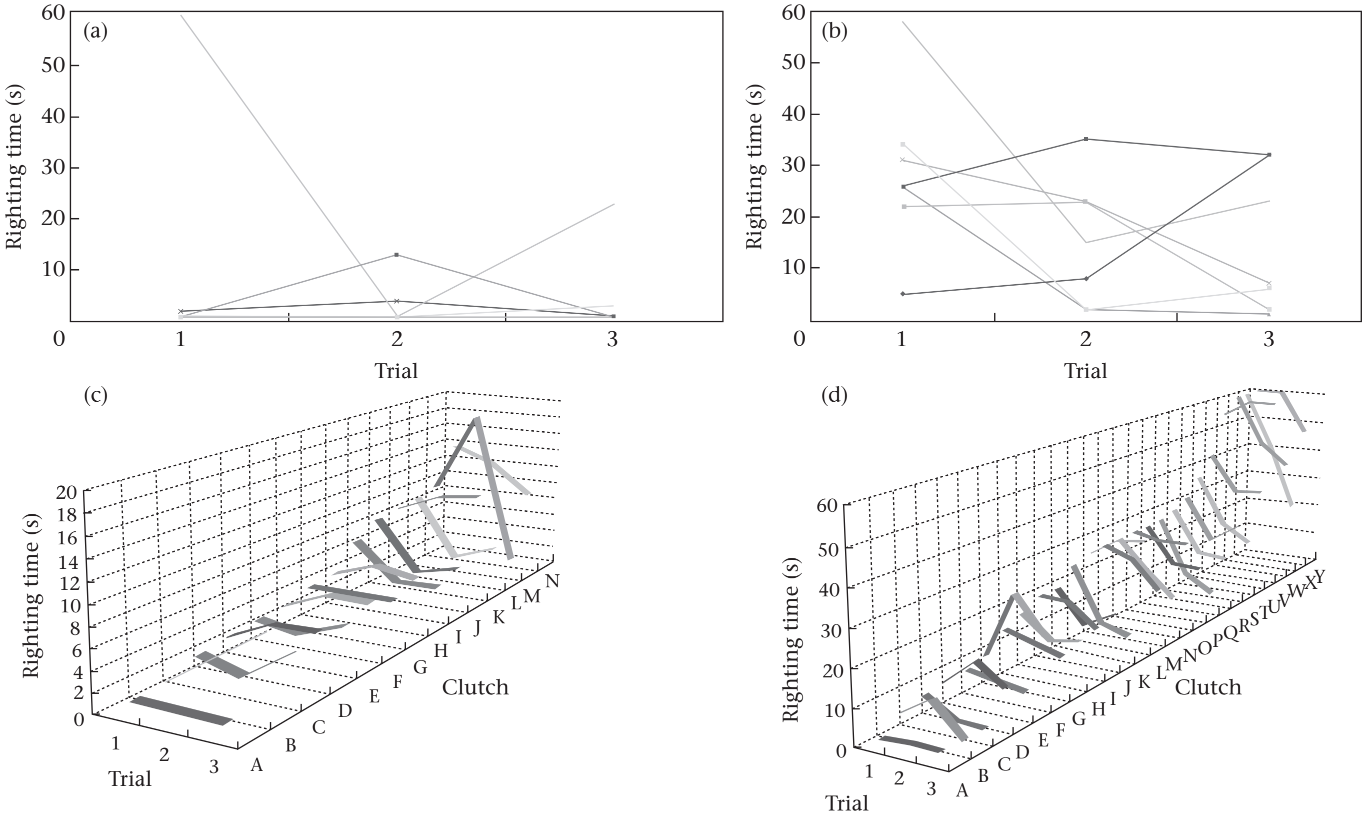

Figure 2. Righting times of individual hatchling turtles in a single clutch of (a) *Apalone spinifera* and (b) *Emydoidea blandingii*. Average righting times for clutches in each trial of (c) *A. spinifera* ($N = 14$) and (d) *E. blandingii* ($N = 25$), ranked from fastest to slowest righting time in trial 1.

Table 2
Rank correlation (Spearman r_S) of clutches based on average righting time of individuals given three consecutive trials (T1–T3)

		T1–T2	T2–T3	T1–T3
Apalone spinifera	r_S	0.55	0.50	0.08
(*N*=14)	*P* value (two-sided)	0.04*	0.07	0.78
Emydoidea blandingii	r_S	0.02	0.90	0.78
(*N*=25)	*P* value (two-sided)	0.91	<0.01†	<0.01†

* Denotes pairs of trials that were significantly correlated at an uncorrected $\alpha = 0.05$.

† Denotes significance after correction for multiple pairwise comparisons.

Furthermore, using our data to test a hypothesis about the relationship between size and righting time resulted in different conclusions depending on which trial was used. Therefore, we reject righting time as an appropriate proxy for relative fitness, performance or 'quality' in hatchling turtles. The concerns raised here apply to any performance measure (or other proxy used to estimate relative fitness or quality) for which rank repeatability cannot be demonstrated.

Righting time is already a complicated proxy for fitness because of variation due to sex, size, experience, endurance and maternal effects (Delmas et al., 2007; Rasmussen & Litzgus, 2010), and our results show that it cannot provide an honest measure of individual performance or fitness because it does not provide replicable ranking of individuals. The lack of significance in our rank correlation tests cannot be ascribed to using criteria that were too stringent (statistically but not biologically representative) for significance, or to type II error, because the majority of trial pairs were not significantly correlated even before correction for multiple pairwise comparisons.

Intraclass correlation coefficients for both species were low (0.165 for *A. spinifera* and 0.246 for *E. blandingii*), also indicating that righting time was not strongly repeatable within individuals and did not explain a significant amount of variation. These values are low compared to ICC values for behaviours such as startling response in anemones (ICC = 0.8–0.9; Briffa & Greenaway, 2011) or boldness in marine mammals (ICC = 0.4, Patrick, Charmantier, & Weimerskirch, 2013). Because repeatability decreases as the time between measurements increases (Bell et al., 2009) and because our three measurements were taken within 5 min, it is likely that studies using multiple trials that are more temporally distinct will suffer from increased variance. Finally, even if ICC values of 0.165 and 0.246 are considered high enough to indicate 'individual repeatability', the problem of a lack of rank repeatability remains.

Table 3
Best generalized linear mixed effects models of righting time (Poisson distribution) for hatchlings of *Apalone spinifera* (*N* = 222) and *Emydoidea blandingii* (*N* = 152) with all turtles sampled three times

Species	Variable	Estimated coefficient	Test statistic*	*P*
Apalone spinifera	Fixed effects			
	Intercept	5.20	*Z*=2.79	0.0053
	Sex(M)	−5.04	*Z*=−2.08	0.037
	SCL	−0.11	*Z*=−2.41	0.016
	Sex(M):SCL	0.11	*Z*=1.98	0.048
	Random effects			
	Individual	−0.90 to 2.19	χ^2_1 = 60.34	<0.0001
	Clutch	−0.13 to 0.62	χ^2_1 = 6.71	<0.01
	Trial	0	χ^2_1 = 0	0.999
	Individual:trial	−1.20 to 2.69	χ^2_1 = 1855.7	<0.0001
Emydoidea blandingii	Fixed effects			
	Intercept	8.36	*Z*=2.95	0.0030
	SCL	−0.18	*Z*=−2.18	0.029
	Random effects			
	Individual	−0.81 to 1.06	χ^2_1 = 130.88	<0.0001
	Clutch	−0.87 to 0.87	χ^2_1 = 7.42	0.006
	Trial	−0.18 to 0.26	χ^2_1 = 16.28	<0.0001
	Individual:trial	−1.89 to 1.60	χ^2_1 = 1621.6	<0.0001

SCL: straight line carapace length.

* Statistical tests are *Z* tests for fixed effects and log-likelihood tests with a chi-square test statistic for random effects. Separate and independent models were fitted for each species.

Conclusions drawn from our repeated tests of the 'bigger is better' hypothesis varied depending on which trial was used in the analysis, demonstrating that intraindividual variation in righting time is strong enough to affect a study's conclusions. In the GLMM, which controlled for clutch, individual and sex, and incorporated all three trials, we detected a significant effect of size on righting time. In this analysis, larger hatchlings of both *A. spinifera* and *E. blandingii* flipped faster than smaller hatchlings. However, the repeatability of this conclusion could not be assessed because we only had one complete set of data with three trials that could not be further partitioned. Thus, we do not consider this result to be more biologically meaningful than the results based on single trials. Previous tests of the 'bigger is better' hypothesis based on righting time have also produced equivocal results. Some found a significant effect of size on righting time (Delmas et al., 2007; Steyermark & Spotila, 2001) while others did not (Micheli-Campbell et al., 2011; Mullins & Janzen, 2006; Paitz et al., 2009; Rasmussen & Litzgus, 2010; Refsnider, 2013). Similarly, size and short-term survival of hatchlings in the wild or in the laboratory were correlated in some studies (Delmas et al., 2007; Janzen, 1993; Janzen, Tucker, & Paukstis, 2000), but not in others (Congdon et al., 1999; Delmas et al., 2007; Paterson, Steinberg, & Litzgus, 2014). Regardless, our goal was not to test the 'bigger is better' hypothesis, but to evaluate whether a particular fitness proxy produced repeatable conclusions when applied to a hypothesis.

We found no effect of incubation temperature (sex) on righting time in *E. blandingii*, in contrast to several previous studies (Delmas et al., 2007; Freedberg et al., 2004; Micheli-Campbell et al., 2011; Mullins & Janzen, 2006; but see Paitz et al., 2009). In contrast, sex had a significant effect on righting time in *A. spinifera*, but this species has genetically determined sex differentiation and all *A. spinifera* eggs were incubated at equal temperatures. This variation in the factors affecting righting time among species and studies further reinforces our caution that this complex behavioural response is not an informative fitness proxy.

Individuals responded differently to repeated trials, as indicated by significant interactions between individual and trial in both *E. blandingii* and *A. spinifera*. Our results support the learning hypothesis (righting time decreases across trials) in *E. blandingii*, but both learning and tiring hypotheses (righting time increases across trials) were rejected in *A. spinifera*. These findings suggest that righting time is further complicated as a fitness proxy by phenotypic variation within species and even within clutches. Behavioural variation among related individuals is an important form of phenotypic plasticity and may increase the adaptive potential of a related cohort born into unpredictable and varying conditions. In turtles, which are iteroparous and long-lived, successive clutches from the same parents will likely experience wide variations in selective pressures due to temporal variation in environmental conditions (Davy et al., 2011). Thus, increased variation (both genetically and environmentally determined) in clutchmates can increase overall survivorship of offspring over the lifetime of a reproductive female.

Conclusions based on ambiguous measures of fitness or performance are problematic not only within academia, but also at a broader scale. For example, results based on righting time and similar performance measures have led to a range of conclusions with implications for research, wildlife conservation and even land

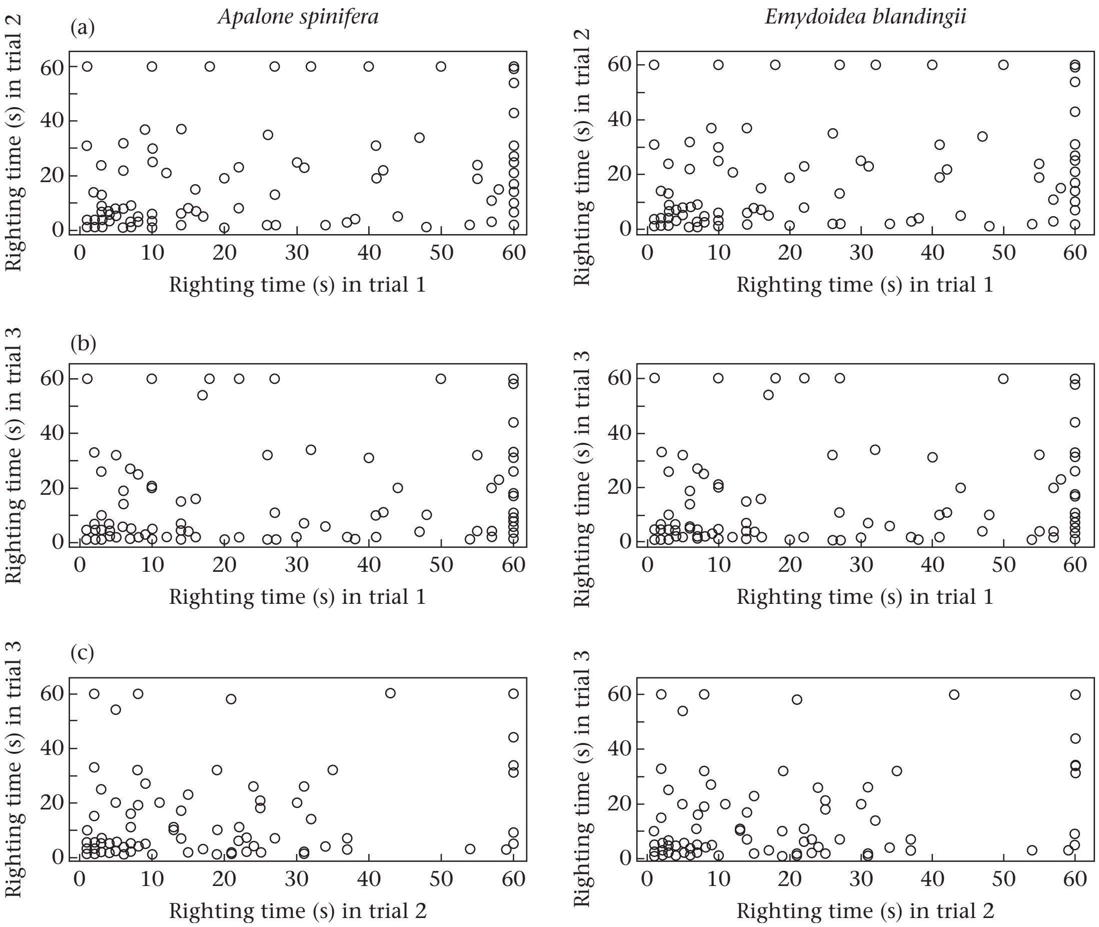

Figure 3. Righting time (in seconds) of individual hatchling *A. spinifera* ($N = 222$) and *E. blandingii* ($N = 152$) for (a) trial 1 versus trial 2, (b) trial 1 versus trial 3, and (c) trial 2 versus trial 3.

management. These include conclusions about the fitness consequences of multiple paternities (Banger et al., 2013), recommendations for hatchery management in threatened sea turtles (Maulany et al., 2012) and the potential effects of predicted climate change on fitness of endangered turtle populations (Micheli-Campbell et al., 2011; Refsnider, Bodensteiner, Reneker, & Janzen, 2013). Thus, the accuracy of performance measures or fitness proxies is not a trivial concern, because the conclusions resulting from their use may have real effects on policy, conservation effectiveness and future research directions.

Conclusions

In summary, our results show that despite a significant individual effect on righting time, righting time displayed low individual and clutch repeatability, and rank order of individuals and clutches varied significantly among trials. We therefore do not consider righting time a reliable proxy for relative fitness or performance. Many behaviours demonstrate repeatability at the individual level (Bell et al., 2009; Dingemanse et al., 2010), but this does not guarantee rank repeatability, which is essential for relative quantities such as fitness and performance (Huey & Dunham, 1987). Other similar measurements (for example, crawling or swimming speed in turtles, or locomotor speed in lizards) may suffer from the same difficulty (Adolph & Pickering, 2008; Huey & Dunham, 1987). A plethora of studies have reiterated the importance of validating fitness proxies (Irschick, Herrel, Vanhooydonck, Huyghe, & Van Damme, 2005; Isden et al., 2013; Wikelski & Romero, 2003; Wilson & Nussey, 2010). Validations of rank repeatability for performance proxies or measures of personality, which are by definition relative and not absolute measurements, are equally critical.

Table 4
Effect of among-trial variation in righting time on tests of the ‘bigger is better’ hypothesis in *Apalone spinifera* ($N = 222$) and *Emydoidea blandingii* ($N = 152$), each given three trials of righting time

Species	Trial	Coefficient of SCL	*t*	*P*	Is bigger better?
A. spinifera	1	−0.08	−1.41	0.159	No
	2	−0.12	−2.35	0.02	Yes
	3	−0.04	−0.72	0.473	No
E. blandingii	1	−0.07	−1.15	0.25	No
	2	−0.12	−1.91	0.06	Possibly
	3	−1.91	−1.99	0.049	Yes

SCL: straight carapace length.

Acknowledgments

This research was funded by a Canada Collections grant from Wildlife Preservation Canada to C.M.D., and a Science Horizons grant from Environment Canada to A.E.L. We thank Ontario Parks for providing logistical support to the project. Pedro Bernardo, Suzanne Coombes, Leif Einarson, Christopher Law, Melissa Oddie and Daniel Storisteanu assisted with nesting surveys, nest collection, hatchling care and righting time trials. We especially thank Mr Ric McArthur for assisting with rapid transport of hatchlings to the release site, ensuring minimal stress to the hatchlings during release. Julia Riley and two anonymous referees provided helpful comments on a previous version of the manuscript.

Supplementary Material

Supplementary material for this article is available, in the online version, at http://dx.doi.org/10.1016/j.anbehav.2014.04.013.

References

Adolph, S. C., & Pickering, T. (2008). Estimating maximum performance: effects of intraindividual variation. *Journal of Experimental Biology, 211*, 1336–1343.

Alford, R. A., Bradfield, K. S., & Richards, S. J. (2007). Ecology: global warming and amphibian losses. *Nature, 447*, E3–E4.

Banger, N., Blouin-Demers, G., Bulté, G., & Lougheed, S. C. (2013). More sires may enhance offspring fitness in northern map turtles (*Graptemys geographica*). *Canadian Journal of Zoology, 91*, 581–588.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42. Retrieved from http://CRAN.R-project.org/package=lme4.

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: a meta-analysis. *Animal Behaviour, 77*, 771–783.

Ben-Ezra, E., Bulté, G., & Blouin-Demers, G. (2008). Are locomotor performances coadapted to preferred basking temperature in the northern map turtle (*Graptemys geographica*)? *Journal of Herpetology, 42*, 322–331.

Bolton, R. M., Marshall, S. A., & Brooks, R. J. (2008). Opportunistic exploitation of turtle eggs by *Tripanurga importuna* (Walker) (Diptera: Sarcophagidae). *Canadian Journal of Zoology, 86*, 151–160.

Bonier, F., Martin, P. R., Moore, I. T., & Wingfield, J. C. (2009). Do baseline glucocorticoids predict fitness? *Trends in Ecology & Evolution, 24*, 634–642.

Bradford, A. L., Weller, D. W., Punt, A. E., Ivashchenko, Y. V., Burdin, A. M., VanBlaricom, G. R., et al. (2012). Leaner leviathans: body condition variation in a critically endangered whale population. *Journal of Mammalogy, 93*, 251–266.

Briffa, M., & Greenaway, J. (2011). High in situ repeatability of behaviour indicates animal personality in the beadlet anemone *Actinia equina* (Cnidaria). *PLoS One, 6*, e21963. http://dx.doi.org/10.1371/journal.pone.0021963.

Burger, J. (1976). Behaviour of hatchling diamondback terrapins (*Malaclemys terrapin*) in the field. *Copeia, 1976*, 742–748.

Burger, J. (1989). Incubation temperature has long-term effects on behaviour of young pine snakes (*Pituophis melanoleucus*). *Behavioral Ecology and Sociobiology, 24*, 201–207.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and multi-model inference: A practical information-theoretic approach*. New York, NY: Springer.

Byrne, P. G., & Robert, J. D. (2000). Does multiple paternity improve fitness of the frog *Crinia georgiana? Evolution, 54*, 968–973.

Congdon, J. D., Nagle, R. D., Dunham, A. E., Beck, C. W., Kinney, O. M., & Yeomans, S. R. (1999). The relationship of body size to survivorship of hatchling snapping turtles (*Chelydra serpentina*): an evaluation of the 'bigger is better' hypothesis. *Oecologia, 121*, 224–235.

Congdon, J. D., Tinkle, D. W., Breitenbach, G. L., & van Loben Sels, R. C. (1983). Nesting ecology and hatching success in the turtle *Emydoidea blandingii*. *Herpetologica, 39*, 417–429.

Davy, C. M., Coombes, S. M., Whitear, A. K., & Mackenzie, A. S. (2010). Visible implant elastomer: a simple, non-harmful method for marking hatchling turtles. *Herpetological Review, 41*, 442–445.

Davy, C. M., Edwards, T., Lathrop, A., Bratton, M., Hagan, M., Nagy, K., et al. (2011). Polyandry and multiple paternities in the threatened Agassiz's desert tortoise, *Gopherus agassizii*: conservation implications. *Conservation Genetics, 12*, 1313–1322.

Delmas, V., Baudry, E., Girondot, M., & Prevot-Julliard, A. (2007). The righting response as a fitness index in freshwater turtles. *Biological Journal of the Linnean Society, 91*, 99–109.

Dingemanse, N. J., Kazem, A. J. N., Réale, D., & Wright, J. (2010). Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution, 25*, 81–89.

Ewert, M. A., & Nelson, C. E. (1991). Sex determination in turtles: diverse patterns and some possible adaptive values. *Copeia, 1991*, 50–69.

Finkler, M. S. (1999). Influence of water availability during incubation on hatchling size, body composition, desiccation tolerance, and terrestrial locomotor performance in the snapping turtle *Chelydra serpentina*. *Physiological and Biochemical Zoology, 72*, 714–722.

Finkler, M. S., & Claussen, D. L. (1997). Use of the tail in terrestrial locomotor activities of juvenile *Chelydra serpentina*. *Copeia, 1997*, 884–887.

Freedberg, S., Stumpf, A. L., Ewert, M. A., & Nelson, C. E. (2004). Developmental environment has long-lasting effects on behavioural performance in two turtles with environmental sex determination. *Evolutionary Ecology Research, 6*, 736–747.

Greenbaum, E., & Carr, J. L. (2001). Sexual differentiation in the spiny softshell turtle (*Apalone spinifera*), a species with genetic sex determination. *Journal of Experimental Zoology, 290*, 190–200.

Huey, R. B., & Dunham, A. E. (1987). Repeatability of locomotor performance in natural populations of the lizard *Sceloporus merriami*. *Evolution, 41*, 1116–1120.

Huey, R. B., & Stevenson, R. D. (1979). Integrating thermal physiology and ecology of ectotherms: a discussion of approaches. *American Zoologist, 19*, 357–366.

Ibáñez, A., Marzal, A., López, P., & Martín, J. (2013). Boldness and body size of male Spanish terrapins affect their responses to chemical cues of familiar and unfamiliar males. *Behavioral Ecology and Sociobiology, 67*, 541–548.

Irschick, D. J., Herrel, A., Vanhooydonck, B., Huyghe, K., & Van Damme, R. (2005). Locomotor compensation creates a mismatch between laboratory and field estimates of escape speed in lizards: a cautionary tale for performance-to-fitness studies. *Evolution, 59*, 1579–1587.

Isden, J., Panayi, C., Dingle, C., & Madden, J. (2013). Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Animal Behaviour, 86*, 829–838.

Jandt, J. M., Bengston, S., Pinter-Wollman, N., Pruitt, J. N., Raine, N. E., Dornhaus, A., et al. (2013). Behavioural syndromes and social insects: personality at multiple levels. *Biological Reviews, 89*, 48–67. http://dx.doi.org/10.1111/brv.12042.

Janzen, F. J. (1993). An experimental analysis of natural selection on body size of hatchling turtles. *Ecology, 74*, 332–341.

Janzen, F. J., Tucker, J. K., & Paukstis, G. L. (2000). Experimental analysis of an early life-history stage: avian predation selects for larger body size of hatchling turtles. *Journal of Evolutionary Biology, 13*, 947–954.

Johnsen, A., Andersen, V., Sunding, C., & Lifjeld, J. T. (2000). Female bluethroats enhance offspring immunocompetence through extra-pair copulations. *Nature, 406*, 296–299.

Kempenaers, B., Verheyena, G. R., & Dhondia, A. A. (1997). Extrapair paternity in the blue tit (*Parus caeruleus*): female choice, male characteristics, and offspring quality. *Behavioral Ecology, 8*, 481–492.

Kingsolver, J., & Huey, R. B. (2008). Size, temperature, and fitness: three rules. *Evolutionary Ecology Research, 10*, 251–268.

Laming, S. R., Jenkins, S. R., & McCarthy, I. D. (2013). Repeatability of escape response performance in the queen scallop, *Aequipecten opercularis*. *Journal of Experimental Biology, 216*, 3264–3272.

Langkilde, T., Lance, V. A., & Shine, R. (2005). Ecological consequences of agonistic interactions in lizards. *Ecology, 86*, 1650–1659.

Lessells, C. M., & Boag, P. T. (1987). Unrepeatable repeatabilities: a common mistake. *Auk, 104*, 116–121.

Maulany, R. I., Booth, D. T., & Baxter, G. S. (2012). The effect of incubation temperature on hatchling quality in the olive ridley turtle, *Lepidochelys olivacea*, from Alas Purwo National Park, East Java, Indonesia: implications for hatchery management. *Marine Biology, 159*, 2651–2661.

Menzies, A. K., Timonin, M. E., McGuire, L. P., & Willis, C. K. R. (2013). Personality variation in little brown bats. *PLoS One, 8*, e80230.

Micheli-Campbell, M. A., Campbell, H. A., Cramp, R. L., Booth, D. T., & Franklin, C. E. (2011). Staying cool, keeping strong: incubation temperature affects performance in a freshwater turtle. *Journal of Zoology, 285*, 266–273.

Mullins, M. A., & Janzen, F. J. (2006). Phenotypic effects of thermal means and variances on smooth softshell turtle (*Apalone mutica*) embryos and hatchlings. *Herpetologica, 62*, 26–36.

Paitz, R. T., Clairardin, S. G., Griffin, A. M., Holgersson, M. C. N., & Bowden, R. M. (2009). Temperature fluctuations affect offspring sex but not morphological, behavioural or immunological traits in the northern painted turtle (*Chrysemys picta*). *Canadian Journal of Zoology, 88*, 479–486.

Parsons, G. R., Foster, D. G., & Osmond, M. (2012). Applying fish behaviour to reduce trawl bycatch: evaluation of the nested cylinder bycatch reduction device. *Marine Technology Society Journal, 46*, 26–33.

Paterson, J.E., Steinberg, B.D, & Litzgus, J.D. (2014). Effects of body size, habitat selection and exposure on hatchling turtle survival. Manuscript submitted for publication.

Patrick, S. C., Charmantier, A., & Weimerskirch, H. (2013). Differences in boldness are repeatable and heritable in a long-lived marine predator. *Ecology and Evolution, 3*, 4291–4299.

Patterson, L. D., & Blouin-Demers, G. (2008). The effect of constant and fluctuating incubation temperatures on the phenotype of black ratsnakes (*Elaphe obsoleta*). *Canadian Journal of Zoology, 86*, 882–889.

Rasmussen, M. L., & Litzgus, J. D. (2010). Patterns of maternal investment in spotted turtles (*Clemmys guttata*): implications of trade-offs, scales of analyses, and incubation substrates. *Écoscience, 17*, 47–58.

Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*, 141–157.

Refsnider, J. (2013). High thermal variance in naturally incubated turtle nests produces faster offspring. *Journal of Ethology, 31*, 85–93.

Refsnider, J. M., Bodensteiner, B. L., Reneker, J. L., & Janzen, F. J. (2013). Nest depth may not compensate for sex ratio skews caused by climate change in turtles. *Animal Conservation, 16*, 481–490.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution, 43*, 223–225.

Riley, J. L., & Litzgus, J. D. (2013). Evaluation of predator-exclusion cages used in turtle conservation: cost analysis and effects on nest environment and proxies of hatchling fitness. *Wildlife Research, 40*, 499–511.

R Development Core Team. (2012). ***R**: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Schwartz, F. J. (1981). A long term internal tag for sea turtles. *Northeast Gulf Science, 5*(1), 87–93.

Seyfarth, R. M., Silk, J. B., & Cheney, D. L. (2012). Variation in personality and fitness in wild female baboons. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 16890–16895.

Shine, R., Langkilde, T., Wall, M., & Mason, R. T. (2005). The fitness correlates of scalation asymmetry in garter snakes *Thamnophis sirtalis parietalis*. *Functional Ecology, 19*, 306–314.

Standing, K. L., Herman, T. B., & Morrison, I. P. (1999). Nesting ecology of Blanding's turtle (*Emydoidea blandingii*) in Nova Scotia, the northeastern limit of the species' range. *Canadian Journal of Zoology, 77*, 1609–1614.

Steyermark, A. C., & Spotila, J. R. (2001). Body temperatures and maternal identify affect snapping turtle (*Chelydra serpentina*) righting response. *Copeia, 2001*, 1050–1057.

Telemeco, R. S., Radder, R. S., Baird, T. A., & Shine, R. (2010). Thermal effects on reptile reproduction: adaptation and phenotypic plasticity in a montane lizard. *Biological Journal of the Linnean Society, 100*, 642–655.

Wikelski, M., & Romero, L. M. (2003). Body size, performance and fitness in Galapagos marine iguanas. *Integrative and Comparative Biology, 43*, 376–386.

Wilson, A. J., & Nussey, D. H. (2010). What is individual quality? An evolutionary perspective. *Trends in Ecology & Evolution, 25*, 207–214.